METHOD FOR OS ACTIVATION OF PLURAL COMPUTERS SHARING STORAGE

Patent number:	JP2001249908
Publication date:	2001-09-14
Inventor:	MIZUNO YOICHI; MATSUNAMI NAOTO; KAMIMAKI HIDEKI; MATSUMOTO JUN
Applicant:	HITACHI LTD
Classification:	
- international:	***G06F1/26; G06F9/445; G06F13/10; G06F13/14; G06F15/177;*** **G06F1/26; G06F9/445; G06F13/10; G06F13/14; G06F15/16;** (IPC1-7): G06F15/177; G06F1/26; G06F9/445; G06F13/10; G06F13/14
- european:	
Application number:	JP20000067150 20000307
Priority number(s):	JP20000067150 20000307

View INPADOC patent family

Abstract of **JP2001249908**

PROBLEM TO BE SOLVED: To solve the problems that access to a storage is concentrated at the time of simultaneously activating plural computers, the processing performance of the storage lowers and long time is required for OS activation in a computer system where the plural computers share the storage.
SOLUTION: At least two or more plural computers, a console for activating the computers and the storage shared by the plural computers are included. The storage is provided with an OS used by the respective computers, the console is provided with a remote power control program for giving the instruction of power ON to the computers and a management table for managing the group constitution of the computers and the computer is provided with a remote power ON control means capable of controlling power supply from a remote location. The remote power control program turns ON the power of the computers for respective groups corresponding to the group constitution of the management table.

JP2001-249908

* NOTICES *
JPO and NCIPI are not responsible for any
damages caused by the use of this translation.
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] In the computer system equipped with the storage which said two or more computers share with two or more at least two or more computers and the console for starting said computer storage It has OS which each of said computer uses. A console It has the managed table which manages the remote power control program which gives directions of power-on to said computer, and the group configuration of said computer. A computer It is the computing system which is equipped with the remote power-on control means which can carry out power control from a remote place, and is characterized by said remote power control program carrying out power-on of said computer for every group according to the group configuration of said managed table.
[Claim 2] The approach of starting of the computer which said remote power control program gives time difference, carries out power-on of said computer for every group in said computer system according to claim 1 according to the group configuration of said managed table, and is characterized by performing OS starting of a computer.
[Claim 3] In the computing system which two or more at least two or more calculating machines and said all calculating machines have connected with the storage in which direct access is possible through the interface for storage Storage is equipped with the local logical volume only prepared in each calculating machines, and a communication link logical volume with all accessible calculating machines. A communication link logical volume It has the calculating-machine status table which records the starting situation of all calculating machines. A calculating machine It has the boot rise control program which controls starting of OS and tells the starting situation of a calculating machine. A calculating-machine status table The starting situation of each calculating machine is indicated by the boot rise control program. A boot rise control program The computing system which enables access distribution to storage and is characterized by performing OS starting of a calculating machine by controlling the timing of OS starting initiation of a calculating machine by the starting situation with reference to a calculating-machine status table in order to know the number of calculating machines under starting.
[Claim 4] The approach of starting of the calculating machine characterized by having the step which said calculating machine carries out power-on, and puts said boot rise control program into operation, the step which said boot rise control program accesses the calculating-machine status table of said communication link logical volume, and refers to the starting situation of a total-session machine, and the step said boot rise control program indicates a starting situation to be to the exclusive field of each calculating machine of said calculating-machine status table in

said computer system according to claim 3.
[Claim 5] In said computer system according to claim 3 storage It has the shared logic volume called a communication link logical volume with all respectively accessible calculating machines and consoles. A communication link logical volume It has the special-purpose-logic field appointed [that only the computer can be only written in each computers, and]. Said special-purpose-logic field The approach of the communication link characterized by recognizing the starting situation in each computer when the starting situation of each computer is indicated and said computer leads said special-purpose-logic field of each computer to suitable timing.
[Claim 6] In the computing system which equipped further the configuration of said computing system according to claim 3 with the console a computer It has the boot rise control program which controls starting of OS and tells the starting situation of a calculating machine. A console It has the computer status table which records the starting situation of a computer starting control program and a computer which controls starting of a computer. A boot rise control program communicates a starting situation to a computer starting control program. A computer starting control program The computing system which makes the temporal dispersion of access to storage possible, and is characterized by performing OS starting of a calculating machine by communicating to a boot rise control program with reference to a calculating-machine status table in order to control the number of startings of a calculating machine.
[Claim 7] The motive approach [calculating machine / which said calculating machine carries out power-on, and is characterized in said computer system according to claim 6 by to have the step which puts said boot rise control program into operation, the step which said boot rise control program notifies that a starting situation is to said calculating-machine starting control program, the step to which said calculating-machine starting control program refers to said calculating-machine status table, the step to which said calculating-machine starting control program rewrites said calculating-machine status table, and the step which said calculating-machine starting control program notifies to said boot rise control program].
[Claim 8] An interface for said console and calculating machine to connect in said computing system according to claim 6 is a fiber channel. A means to control the fiber channel with which said console and calculating machine are equipped with a means to control a fiber channel, and said console and calculating machine are equipped World Wide Name which is the message frame which can be broadcast and is an identifier for specifying a means to control the fiber channel mounted in at least one or more calculating machines as the interior, It is ready-for-sending ability about the message frame which stored the starting control identifier for specifying the control information which the starting situation and said console of said computer publish. A means to control said fiber channel, on the message frame of broadcasting which said console and calculating machine published The approach of a communication link of the computer characterized by performing starting control according to the contents of said starting control identifier when it checks that World Wide Name which is an identifier for specifying a means to control the fiber channel concerned is stored, and a console.
[Claim 9] It is the approach of starting of the computer characterized by displaying that on the output unit of said computer in order that said boot rise control program may tell the operator of said computer that it is in an waiting condition by the number control of startings in a computer system said claim 3 and given in six.
[Claim 10] The boot rise control program with which said calculating machine is equipped in a computer system said claim 3 and given in six is a computer system characterized by being the program performed at the time of power-source starting, having controlled the number of

startings of said calculating machine, and having a boot rise control program rewritable by said calculating machine.
[Claim 11] It is the approach of starting of the computer system characterized by the interface for storage being a fiber channel in a computer system given in ten or the approach of starting of a calculating machine from said claim 1 and a calculating machine.
DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention] This invention relates to the approach of preventing the degradation by the storage centralized access of two or more calculating machines especially at the time of OS starting, about the approach of OS starting of two or more calculating machines which share storage.
[0002]
[Description of the Prior Art] The computer system equipped with the server which manages two or more calculating machines and two or more calculating machines as a computer system using two or more calculating machines, and the storage which two or more calculating machines share is known. It explains using the example of drawing 17 .
[0003] The storage with which a calculating machine shares 200x (200a=200n), and calculating-machine 200x share 100 in this drawing, OS to which computer 200x download 1001, the server which 400 has in the location distant physically [the location in which the computer is installed], The connection interface to which 500 connects storage 100 with a server 400, and 600x (600a=600n) are LANs (Local Area Network) for two or more calculating-machine 200x and servers 400 to communicate.
[0004] If computer 200x carry out power-on, it is computer 200x first. A server 400 is accessed by LAN 600x course. Next, OS 1001 in storage 100 are downloaded and starting is started. Two or more computers All computers that carried out power-on when 200x carried out power-on It is a server in order that 200x may download OS 1001. The load to 400 or storage 100 will increase.
[0005]
[Problem(s) to be Solved by the Invention] (Technical problem) With the above-mentioned conventional technique, when it is going to start two or more calculating machines to coincidence, access will concentrate on a server or storage, in the processing which a load like OS starting requires, the processing engine performance falls, and the technical problem that OS starting takes great time amount occurs. When OS warm-up time is made huge and uses the computer of the many number, it will cause trouble to business, so that there are many computers started to coincidence.
[0006] If each operator shifts and starts between power ups, the access concentration to storage is avoidable, but it grasps when other operators switch on a power source, and about starting of an alien machine, always taking between power ups into consideration, if it kicks, waiting and the technical problem which switches on a power source that it does not become are in a computer. Whenever powering on uses a computer, he has to carry out, and he has the technical problem that an operator takes excessive time and effort each time.
[0007] (Purpose) The purpose of this invention is preventing the degradation by the access concentration to storage, and shortening OS warm-up time in the computing system which shares storage by offering the approach of offering the approach of two or more computers starting, without a manager or an operator spending the new time and effort and the time amount about

starting of a computer, and distributing the load to storage in time at the time of OS starting of two or more computers.
[0008]
[Means for Solving the Problem] In order to attain the above-mentioned purpose, the computing system equipped with the storage which a management console, and two or more computers and management consoles for carrying out power-on of two or more computers and two or more computers share is built.
[0009] A management console is equipped with the managed table for managing the remote power control program and computer group for controlling the power source of two or more computers from a remote place with the first means of this invention.
[0010] A computer is equipped with the remote power-on control means which can carry out power control from a remote place.
[0011] A remote power control program makes the temporal dispersion of access to storage possible by giving and carrying out power-on of the time difference for every computer group according to the computer group defined by the managed table.
[0012] With the second means of this invention, two or more calculating machines and management consoles are connected to storage with connection interfaces, such as a fiber channel.
[0013] A management console is equipped with the remote power control program for controlling the power source of two or more computers from a remote place.
[0014] A computer controls the remote power-on control means which can carry out power control from a remote place, and starting of OS, and is equipped with the boot rise control program which tells the starting situation of a computer.
[0015] Storage is equipped with the Logical unit of respectively dedication to two or more computers, the communication link Logical unit which can use all computers and records the starting situation of a computer, and LU definition program which defines and creates such Logical unit.
[0016] Access distribution to storage is enabled and OS starting of a calculating machine is performed because a boot rise control program controls the number of calculating machines with reference to communication link Logical unit.
[0017] Moreover, in said second means, the third means of this invention equips a management console with the computer starting control program which controls starting of a computer instead of the communication link Logical unit arranged at storage, and is characterized by performing concentration mold starting control.
[0018]
[Embodiment of the Invention] (The first operation gestalt) Although the system configuration which used the server was taken in the Prior art, it considers as the system configuration of a server loess mold excellent in the field of operational administration or the engine performance with this operation gestalt. The technical problem that the access concentration to storage takes place at the time of starting of two or more calculating machines which were stated with the conventional technique also in this case is the same. In addition, of course, this invention is not limited to this configuration.
[0019] Drawing 1 is the block diagram of the computing system of this operation gestalt. In this drawing, a management console for the storage with which a calculating machine shares 2x (2a=2n), and all the calculating machines 2 share 1, and 4 to manage a computer system, the fiber channel connecting means by which 3 connects all the calculating-machine 2x, storage 1,

and management consoles 4 mutually, and 5x (5a=5n) are fiber channels. 7 is storage 1 and a management console. Means of communications for 4 to communicate and 6x (6a=6n) are LANs (Local Area Network) for two or more computer 2x and management consoles 4 to communicate.

[0020] Hereafter, although explained as a computer supposing computers for clients, such as PC (Personal Computer), it is applicable also to a server and its alien machines, such as a workstation. Moreover, although the interface (it is written as I/F below) for connecting storage with a calculating machine assumes the optimal fiber channel in respect of the connection distance and rate, it is applicable to various I/F, such as SCSI (Small Computer System Interface), USB (Universal Serial Bus), and IEEE1394.

[0021] Drawing 2 is the block diagram of computer 2x. In order that an I/O device for the user of this computer 2 to input or output information, as for 21, a CC means to by_which 22 controls the whole computer, and 25 may perform connection of a fiber channel and control, the fiber channel board which carried in a computer 2, and 24 are memory for the LAN board for performing the used communication link and 23 to store a program and data required in order that the CC means 22 may perform various control LAN 6 x.

[0022] In the fiber channel board 25, the fiber channel controller by which 250 controls a fiber channel, and 251 are the memory for storing a program and data required in order that a fiber channel controller may control, a program required in order that the CC means 22 may control the fiber channel controller 250, and data. In memory 251, 2511 is a boot rise control program which the CC means 22 performs, in order to control the fiber channel controller 250 to ** which starts OS from the storage 1 of the exterior of calculating-machine 2x linked to fiber channel 5x. 2512 is WWN (World Wide Name) which is the information on a name only in the world established in order to identify the fiber channel controller 250 uniquely. Although [here / this operation gestalt] the fiber channel board 25 is carried in calculating-machine 2x, it is also possible to mount directly the fiber channel controller 250 carried in this board and memory 251, without using this board 25 for calculating-machine 2x, and an operation of this invention and effectiveness are completely the same also in this case.

[0023] 26 is the Maine electric power switch of calculating-machine 2x.

[0024] Drawing 3 is the block diagram of storage 1. Means of communications for memory for a CC means by which 11 manages control by the whole storage, the fiber channel controller, to which 15 performs the connection and its control to fiber channel 5d, and 12 to store a program and data for the CC means 11 to perform internal control of storage 1, and 16 to perform the communication link between the management consoles 4, and 13x (13a=13n) and 14 are Logical unit (LU:Logical Unit).

[0025] Logical unit LU is the imagination volume prepared in the storage 1 interior called a logical volume, and is the name defined in the specification of SCSI (Small Computer System Interface) which is one protocol of I/F which connects storage with a calculating machine here. Hereafter, the thing of a logical volume will only be called LU. Moreover, the thing of the number for identifying LU is called LUN (Logical Unit Number).

[0026] Logical unit 13x are LU of computer 2x dedication, and are local LU for computer #x. OS 131x only for [computer 2x] each are stored in local LU13x. on the other hand -- Logical unit 14 -- any of all the computer 2x and management consoles 4 -- although -- it is the accessible common use LU.

[0027] LU definition program which the CC means 11 performs in memory 12 in order that 121 may define or create two or more LUs, 122 -- each -- access prohibition / authorization of LU of

plurality [x / computer 2] which is alike, respectively and receives -- A LUN managed table for the CC means 11 to manage each correspondence relation of the interior LUN added in order to manage LU inside the imagination LUN which is the attribute of LU, and LUN recognized from a host computer, and storage, 123 -- the information on the LUN managed table 122 -- following -- each -- when a permission is granted by performing control to which access to each LU from computer 2x is restricted / permitted, it is an access-control program for the CC means 11 to perform control of read/write access to the LU concerned.

[0028] LU is explained. LU is the logical volume when seeing from the host computer of storage. A host computer recognizes one LU as one set of a logical disk unit.

[0029] Storage 1 can define and build two or more LUs inside. This will be called Interior LU. At storage 1, in order to manage Interior LU, serial number attachment is carried out for the integer which begins from 0. This number is called Interior LUN.

[0030] the case where two or more calculating-machine 2x, on the other hand, share one set of storage 1 like the computer system of this invention -- calculating-machine 2x -- it is alike, respectively and LU of dedication is assigned. Generally, although host computers, such as PC, search the storage connected at the time of OS boot and LU is detected, the search approach may have some constraint. It is (a). LUN is (b) searched sequentially from 0. When it assumes that LUN exists with the consecutive number and a certain number does not exist, they are two not performing subsequent searches. This is a device for shortening the search time. It is assumed that it is computer 2x in which the computer 2 of this invention also has such a property. In such a case, supposing it assigns Interior LUN to a host computer as it is, the computer which was able to assign LUs other than internal LUN=0 can detect this LU. Then, it is desirable to begin from 0 and to assign LUN with the consecutive number to all computers. Then, in this invention, storage 1 solves the above-mentioned technical problem by redefining all internal LU ** that the computer 2x use for each computer 2x of every so that it may be set to continuous LUN which begins from 0. thus -- each -- LU recognized from computer 2x is distinguished from Imagination LU, and the number is distinguished from the interior [Imagination / LUN /, call, and LUN] LU. Section LUN and the correspondence relation with Imagination LUN are managed among these on the LUN managed table 122 with which storage 1 is equipped.

[0031] An example of the LUN managed table 122 with which storage 1 is equipped is shown in drawing 4. An attribute is stored in the LUN managed table 122 with a port number, Target ID, Imagination LUN and Interior LUN and WWN, and S_ID.

[0032] A port number stores the number of the fiber channel connection port with which storage 1 is equipped. With this operation gestalt, the number of ports assumes one piece and stores 0 uniformly.

[0033] TargetID is the discernment ID assigned to storage 1 in connection I/F of calculating-machine 2x. Although it has only D_ID (Destination ID) for every port when connection I/F of calculating-machine 2x and storage 1 is a fiber channel like this operation gestalt, since there is a term of a port number, you may omit, and D_ID determined at the time of initialization of a fiber channel may be stored. Since it can have two or more ID in the same port in the case of SCSI, Target ID to which each LUN at that time belongs is stored. Since the fiber channel is assumed with this operation gestalt, the column of Target ID presupposes that it is intact, and stores 0 uniformly.

[0034] the interior LUN which assigned Imagination LUN and Interior LUN to computer 2x -- each -- correspondence relation with the imagination LUN recognized from computer 2x is shown. For example, the interior LUN 0 defines the interior LUN 2 as imagination LUN 0 as

imagination LUN 0. Although both both are defined as imagination LUN 0, computer 2x which can be used, respectively differ.

[0035] WWN -- each -- World Wide Name which is the information which specifies the fiber channel controller 250 of calculating-machine 2x is stored. the time of port log in processing in which the connection relation between the port of a fiber channel and a port is established -- each -- WWN 2512 of calculating-machine 2x is notified to storage 1.

[0036] S_ID is ID information stored in the frame header of a fiber channel, and is ID which identifies the source (initiator) which created the frame. S_ID is dynamically assigned in the case of initialization of a fiber channel. Although WWN described previously is a value set as a meaning by each fiber channel port where it was exchanged on the occasion of initialization, even if it does not investigate WWN for every frame by performing related attachment of WWN and S_ID, computer 2x can be specified by inspecting only S_ID.

[0037] An attribute shows the possession attribute of each LU. It is shown that "dedication" is LU of one computer 2x dedication. It is shown that "common use" is LU which two or more computer 2x share.

[0038] Here, the correspondence relation between the block diagram of the storage 1 of drawing 3 and the LUN managed table 122 of drawing 4 R> 4 is explained. WWN=WWNa, S_ID=S_IDa, and computer 2b WWN=WWNb, [computer 2a] If WWN=WWNn, S_ID=S_IDn, and the management console 4 consider as WWN=WWNz and S_ID=S_IDz, S_ID=S_IDb and 2n of computers In local LU13a for computer 2a of drawing 3, internal LUN=0, virtual LUN=0, and local LU13b for computer 2b internal LUN=1 and local LU13n for virtual LUN= 0 and 2n of computers Internal LUN=n, Virtual LUN=0 and computer 2a, 2b, ..., the 2n common use LU 14 map with internal LUN=k and virtual LUN=1. Moreover, since common use LU 14 is shared also by the management console 4, WWN=WWNz and internal LUN=k corresponding to S_ID=S_IDz are mapped as virtual LUN=0. In addition, since the management console 4 does not possess local LU of dedication in storage 1 here, common use LU 14 is mapped as virtual LUN=0.

[0039] The operator who operates the management console 4 can send the directions which define and create LU from I/O device 41 of the management console 4. These directions are published by storage 1 by means of communications 47. The means of communications 16 of storage 1 receives these directions, and reports them to the CC means 11. The CC means 11 starts LU definition program 121, and creates LU according to directions. And an attribute is stored in the LUN managed table 122 as access-control information on defined LU with a port number, Target ID, Imagination LUN and Interior LUN and WWN, and S_ID. LU can be created as the above-mentioned LUN managed table 122 by repeating the number of LUs which assign this activity to all calculating-machine 2x.

[0040] Drawing 5 is the block diagram of the management console 4. An I/O device for a manager to operate the management console 4, as for 41, a CC means by which 42 manages control of the management console 4 whole, Memory for 43 to store the program and data which the CC means 41 performs, A LAN board for the fiber channel board on which 45 performs connection with fiber channel 5e and control, and 44 to perform the communication link which used LAN6d, the disk unit in which 46 stores OS and the program of the management console 4, and 47 are means of communications which perform the communication link between storage.

[0041] In memory 43, the remote power control program with which 431 controls the power source of computer 2x, and 4311 are managed tables which manage a computer group. The configuration of this table 4311 is shown in drawing 6 . As for this table 4311, the group number

of each computer is indicated. The calculating machine belonging to the same group will be started to coincidence, and the engine performance of storage etc. determines the number of calculating machines per group.

[0042] This drawing is an example when setting the number of coincidence startings to 3. In addition, if there will be neither modification of a computer configuration nor modification of the processing engine performance of storage once it creates this table 4311, it can be used as it is also from next time.

[0043] Next, detail actuation of starting of a computer is explained.

[0044] (1) The operator who operates the managed table creation management console 4 divides two or more computers to start into some groups, and registers the group number of each computer into the managed table 4311 from I/O device 41. The group number is set to 0 about the computer which is not started.

[0045] (2) According to the managed table 4311, the power-on remote power control program 431 of a calculating machine performs creation and transmission of the remote power-on packet 900 at fixed spacing for every group sequentially from the group number 1, and is a calculating machine. 2x are started. Let preferably sending-out spacing of the remote power-on packet 900 be the warm-up time of a computer, and abbreviation identitas. This packet 900 is a packet for carrying out power-on of computer 2x, and shows the configuration to drawing 7. This packet 900 consists of a header 901 and data division 902. Data division 902 are what continued n times, and MAC Address 9022 of computer 2x which carry out remote power-on to a synchronizer 9021 is constituted.

[0046] The remote power control program 431 controls the LAN board 44, and transmits this packet 900 to computer 2x from this board 44.

[0047] The remote power-on control means 241 on the calculating-machine 2 LAN board 24 of x receives this packet 900. This means 241 detects MAC Address 9021 in the data division 902 of this packet 900. This means 241 compares detected MAC Address 9021 with the MAC Address of the LAN board 24, when the same, it controls the Maine electric power switch 26, and it carries out power-on of computer 2x. Calculating-machine 2x which carried out power-on read and start OS 131x only for [in local LU13x of storage 1 / calculating-machine 2x] each.

[0048] Thus, calculating-machine 2x give time difference for every group, and since they carry out sequential starting, prevention of the access concentration to storage of them is attained.

[0049] Since according to this operation gestalt starting of a computer is performed for every computer group and a fixed number of computers start, access to storage can be distributed in time and superfluous access concentration can be controlled. Therefore, the degradation of storage can be prevented and OS warm-up time in the whole system can be shortened.

[0050] (The second operation gestalt) With the first operation gestalt, in the case of the planned power-on from the remoteness by the management console, it is effective, but in the case of the direct power-on by the operator, the technical problem that it cannot respond remains. Therefore, in the environment where both the power-on by the management console and the power-on by the operator are intermingled, access may concentrate to storage. The access concentration preventive measure which is not depended is required for the approach of power-on so that it can respond also in the environment where the approach of power-on is intermingled. This operation gestalt can prevent the access concentration to storage, without being based on the approach of power-on, and can start two or more computers. Here, the power-on of the computer from remoteness is explained supposing being a LAN course.

[0051] Drawing 8 is the block diagram of the computing system of this operation gestalt. The

difference with the first operation gestalt is storage. It is having newly equipped 1 with communication link LU 17. communication link LU 17 -- any of all the computer 2x and management consoles 4 -- although -- it is accessible and is LU for communicating the starting situation of a computer. The calculating-machine status table 171 is stored in communication link LU 17. This table 171 is a table showing the starting situation of calculating-machine 2x.
[0052] The configuration of this table 171 is shown in drawing 9. The flag block for controlling exclusively and the block for calculating-machine 2x are shown in this table 171, and the address in this table 171 is decided to be a meaning, respectively. The status 1 of a flag block means that one of calculating-machine 2x is accessing a table, and the status 0 means that the calculating machine under access on a table is not. It means that the calculating machine is OS starting the status 1 of the block for calculating-machine 2x, and means that un-starting ended the status 0 and starting ended the status 2. Not the file read/write with which the boot rise control program 2511 of calculating-machine 2x used the file system but RAW which carries out read/write of the logical block of direct LU It accesses by the I/O command. moreover, the thing which the management console 4 leads this calculating-machine status table 171 by RAW I/O, and is polled to suitable timing -- each -- the situation of calculating-machine 2x can be known serially.
[0053] The boot rise control program 2511 is used in order to start OS from the storage 1 of the exterior of calculating-machine 2x linked to fiber channel 5x or to communicate with communication link LU 117. Rewriting of this program 2511 can also be enabled by the computer 2 by using a flash memory etc.
[0054] Next, it explains, using the flow chart of drawing 10 about detail actuation of starting of a calculating machine.
[0055] (a) Computer 2x carry out power-on by the power-on from the remoteness by the remote power control program 431 of the power-on management console 4 of a computer, the direct power-on by the operator, etc. (step 701 of drawing 10).
[0056] (b) If power-on of boot rise control program starting calculating-machine 2x is carried out, they will put the boot rise control program 2511 in the memory 251 of the fiber channel board 25 into operation (step 702).
[0057] (c) Refer to the calculating-machine status table 171 which the reference boot rise control program 2511 has in the communication link LU 17 of storage 1 in a calculating-machine status table for a boot rise control program.
[0058] This program 2511 accesses a flag block first, and investigates whether the computer under access is. If the calculating machine under access is not, a flag block is rewritten to 1, and the block for calculating-machine 2x is referred to. If an alien machine is accessing, it will stand by till access termination (step 703).
[0059] (d) The boot rise control program 2511 whose number of calculating machines under starting is below a threshold can know the number of calculating machines under starting by the total number of the status 1 of the block for calculating-machine 2x of the calculating-machine status table 171. With the engine performance of storage etc., a threshold is a value determined beforehand and expresses the number of computers in which coincidence starting is possible (step 704).
[0060] (e) When there are more calculating machines under screen-display starting than a threshold, the status of a flag block is rewritten for a standby condition to 0, and once end access, and an alien machine ends starting of OS, and the boot rise control program 2511 stands by until the number of calculating machines under starting becomes below a threshold. In order to tell an operator that this program 2511 is in a standby condition in that case, it indicates that it is in a

standby condition on the screen of I/O device 21. This program 2511 accesses the calculating-machine status table 171 again, after carrying out fixed time amount progress (step 709).
[0061] (f) With [the number of calculating machines under starting a boot rise control program rewrites / starting / the status of a self-calculating machine to 1] a threshold [below], the boot rise control program 2511 progresses to the next processing. Since it expresses that it is under starting, this program 2511 rewrites the status of the self-calculating machine of the calculating-machine status table 171 to 1, returns the status of a flag block to 0, and ends access to this table 171 (step 705).
[0062] (g) A boot rise control program starts starting of OS 131x which OS starting initiation boot rise control program 2511 has in local LU 13only for computers x of storage 1 (step 706).
[0063] (h) A boot rise control program rewrites the status of a self-computer to 0 after the completion of OS starting.
[0064] Since it means that starting of OS 131x was completed, the boot rise control program 2511 rewrites the status of the self-calculating machine of the calculating-machine status table 171 to 0 (step 707).
[0065] Starting actuation of a computer is completed above. When two or more computer 2x perform starting actuation to coincidence, starting processing of OS 131x is performed to coincidence to a threshold, but when a threshold is exceeded, alien-machine 2x will stand by until computer 2x under starting complete starting of OS 131x and rewrite the status of the computer status table 171. By this, it is lost that the calculating machine more than a certain fixed numbers performs OS starting processing to coincidence, and the superfluous access concentration to storage 1 can be controlled. Therefore, the degradation of storage 1 can be prevented and OS warm-up time in the whole system can be shortened.
[0066] According to this operation gestalt, at the time of OS starting of two or more calculating machines, the load to storage can be distributed in time and the superfluous access concentration to storage can be controlled in the computer system which shares storage. Therefore, the degradation of storage can be prevented and OS warm-up time in the whole system can be shortened.
[0067] Moreover, since it cannot be based on the approach of power-on but can apply, also in the environment where the power-on and the direct power-on from remoteness are intermingled, it is effective.
[0068] Moreover, starting of two or more computers is attained without a manager or an operator spending the new time and effort and the time amount about starting of a computer, and it leads to reduction of management cost or labor costs.
[0069] (Operation gestalt 3) Starting of the computer which distributed access to storage, without being able to apply also in this operation gestalt not only the power-on of the computer from the remoteness by the management console but in the case of the direct power-on by the operator, and being based on the approach of power-on is possible. Here, it explains supposing the power-on of the computer via the fiber channel by remote operation.
[0070] Drawing 11 is the block diagram of the computing system of this operation gestalt. It is the difference with the second operation gestalt having arranged LAN 6 of drawing 8 not existing, and the calculating-machine status table 171 arranged communication link LU 17 in the management console 4, and having arranged the calculating-machine starting control program 432 newly, and others are the same.
[0071] Drawing 12 is the block diagram of the management console of this operation gestalt. The differences with the second operation gestalt are a LAN board's not existing and having formed

the calculating-machine starting control program 432 newly, and having stored the calculating-machine status table 171 in the disk unit 46. The computer starting control program 432 is a program for controlling the number of computers which carries out OS starting processing. The calculating-machine status table 171 is the same as that of what is shown in drawing 9, and it is a table showing the starting situation of a calculating machine. Although the boot rise control program 2511 carried out the direct reference and read/write was performed with the second operation gestalt, with this operation gestalt, the boot rise control program 2511 is notified to the computer starting control program 432, and the computer starting control program 432 performs read/write.

[0072] Drawing 13 is the block diagram of computer 2x of this operation gestalt. The differences of calculating-machine 2x of the second operation gestalt are that a LAN board does not exist and that the remote power-on control means 2501 and the standby power source 2502 exist on a fiber channel board.

[0073] The standby power source 2502 is formed in computer 2x, and as turned on, it constitutes only this standby power source 2502 also from a condition that the Maine power source of computer 2x is off.

[0074] The remote power-on control means 2501 was formed in the interior of the fiber channel controller 250, and is connected to the standby power source 2502. Therefore, also in the state of OFF of the Maine power source of calculating-machine 2x, it is controllable so that the fiber channel controller 250 operates. Then, the frame (packet) which arrives by fiber channel 5 course is receivable.

[0075] Next, it explains, using the flow chart of drawing 14 about detail actuation of starting of a calculating machine.

[0076] (a) Explain the remote power-on control by the power-on fiber channel of a calculating machine. Drawing 15 is 1 operation gestalt of the remote power-on frame 950 which used the fiber channel. 9501 is SOF (Start of Frame) and shows the head of a frame. 9502 is a frame header and stores information, such as information on the transmitting node of a frame, and classification of a frame. It is shown that 9502a is a broadcasting identifier and is a frame with which this frame is published by coincidence to all the nodes linked to the same fiber channel as a broadcasting frame. 9503 is a payload and is the actual condition of the transfer data of this frame. 9503x (x is a=n) are WWN, and store WWN 2512 of the fiber channel controller 250 of calculating-machine 2x which carry out remote power-on. 9504 is CRC (Cyclic Redundancy Code) and is a data guarantee code for detecting bit transformation of a frame. It is shown that 9505 is EOF (End Of Frame) and is the tail of a frame.

[0077] Next, actuation of the remote power-on control using a fiber channel is explained. The remote power control program 4311 of the management console 4 creates the remote power-on frame 950. At least one or more sets of WWN(s) of computer 2x set as the object of remote power-on at this time are stored in the payload 9503 of a frame 950. And the fiber channel board 45 is controlled and this frame 950 is transmitted by broadcasting.

[0078] here -- each -- it is assumed that all computer 2x have stopped. Although calculating-machine 2x body has stopped, the fiber channel controller 250 is operating with the standby power source 2502, and is awaiting the frame which arrives by the fiber channel 5. Here, supposing it receives the above-mentioned remote power-on frame 950, the remote power-on control means 2501 of this controller 250 will inspect the frame header 9502 of this frame 950, and it will check that it is a broadcasting frame. Next, it inspects whether WWN2512 of this controller 250 is stored in the payload 9503 of this frame 950. If stored, this means 2501 will be

made power-source ON by controlling the Maine electric power switch 26, and will start computer 2x. If not stored, this frame 950 is disregarded, and it awaits again, and will be in a condition.

[0079] Here, WWN stored in the payload 9503 of the remote power-on frame 950 may repeat the same WWN in the first operation gestalt like the remote power-on packet by LAN of a publication, and may store it in it several times.

[0080] As mentioned above, even if there is no LAN, remote power-on control is attained using the fiber channel 5.

[0081] In the above, although how to perform power-on control using the broadcasting frame of a fiber channel was explained, supposing the fiber channel controller supports IP protocol, there is also an approach using the remote power-on packet shown in drawing 7. Also in this case, the above, a configuration, and actuation are the same. However, since a packet can be used as it is, a remote power control program is applicable as it is.

[0082] Thus, power-on of two or more computers 2x is carried out (step 801 of drawing 14).

[0083] (b) If power-on of boot rise control program starting calculating-machine 2x is carried out, they will put the boot rise control program 2511 in the memory 251 of the fiber channel board 25 into operation (step 802).

[0084] (c) Notify a boot rise control program to the calculating-machine starting control program of a management console, and the boot rise control program 2511 which waits for starting authorization sends out the starting control frame 960 in order to notify that it is starting initiation to the calculating-machine starting control program 432 of the management console 4. The starting control frame 960 is a frame used in order that the boot rise control program 2511 may notify starting initiation and termination of computer 2x to the computer starting control program 432 or the computer starting control program 432 may give starting authorization to the boot rise control program 2511. The configuration of the starting control frame 960 is shown in drawing 16. The difference with the remote power-on frame 950 is that the starting identifier 96032 exists. The computer starting control program 432 can specify a computer by WWN 96031, and the computer can know starting initiation or starting termination by the starting identifier 96032 (step 803).

[0085] (d) In order to know the number of calculating machines under starting, refer to the calculating-machine status table 171 in a disk unit 46 for the calculating-machine starting control program 432 with which the calculating-machine starting control program received the notice of reference for the calculating-machine status table. This table 171 is the same as that of what is shown in drawing 9 R> 9, and is a table showing the starting situation of calculating-machine 2x. In addition, with this operation gestalt, only a management console can refer to this table 171, and since exclusive control is unnecessary, a flag block is not used (step 804).

[0086] (e) The calculating-machine starting control program 432 whose number of calculating machines under starting is below a threshold can know the number of calculating machines under starting with the number of the statuses 1 of the calculating-machine status table 171. The notice of starting termination comes from an alien machine, and the boot rise control program 2511 is made to stand by, when there are more computers under starting than a threshold until the number of computers under starting becomes below a threshold. At this time, the calculating-machine starting control program 432 sends out the starting control frame 960 in order to transmit a standby instruction to the boot rise control program 2511. With [the number of computers under starting] a threshold [below], the computer starting control program 432 progresses to the next processing (step 805).

[0087] (f) If the screen-display starting control frame 960 is sent out in a standby condition, it can know that the boot rise control program 2511 judged whether it was a frame addressed to itself by WWN96031, and received the standby instruction by the starting control identifier 96032. This program 2511 that received the standby instruction indicates that it is in a standby condition on the screen of I/O device 21 in order to tell an operator that it is in a standby condition. This program 2511 sends out the starting control frame 960 of starting initiation to the computer starting control program 432 of the management console 4 again, after carrying out fixed time amount progress (step 812).

[0088] (g) With [the number of computers under starting a computer starting control program rewrites / starting / to 1 the status of the computer which received the notice] a threshold [below], the computer starting control program 432 progresses to the next processing. Since it means that the calculating machine which received the notice is starting, this program 432 rewrites to 1 the status of the calculating machine which received the notice of the calculating-machine status table 171 (step 806).

[0089] (h) The calculating-machine starting control program 432 with which a calculating-machine starting control program gives starting authorization to a boot rise control program sends out the starting control frame 960 in order to give starting authorization to the boot rise control program 2511. It can know that the boot rise control program 2511 which was standing by judged whether it was a frame addressed to itself by WWN 96031, and obtained starting authorization by the starting control identifier 96032 (step 807).

[0090] (i) The boot rise control program 2511 with which the boot rise control program which obtained starting authorization obtained OS starting initiation starting authorization starts starting of OS 131x in local LU13only for computers x of storage 1 (step 808).

[0091] (j) After the completion of OS starting, a boot rise control program sends out the starting control frame 960 to it, in order that the notice boot rise control program 2511 of the completion of starting may notify that starting of OS 131x was completed to the calculating-machine starting control program 432 to a calculating-machine starting control program (step 809).

[0092] (k) Since it means that starting of OS 131x which rewrite to 0 the status of the computer which received the notice completed the computer starting control program, the computer starting control program 432 rewrites to 0 the status of the computer which received the notice of the computer status table 171 (step 810).

[0093] Starting actuation of a computer is completed above. When two or more computer 2x perform starting actuation to coincidence, starting processing of OS 131x is performed to coincidence to a threshold, but when a threshold is exceeded, alien-machine 2x will stand by until computer 2x under starting complete starting of OS 131x and notify starting termination to a management console. By this, it is lost that the calculating machine more than a certain fixed numbers performs OS starting processing to coincidence, and the superfluous access concentration to storage 1 can be controlled. Therefore, the degradation of storage 1 can be prevented and OS warm-up time in the whole system can be shortened. This operation gestalt can be called number control system of concentration mold startings by the management console to the second operation gestalt having been the number control system of distributed startings by each computer.

[0094] Since according to this operation gestalt the above passage remote power-on control can be performed via a fiber channel with two or more computers and a communication link required between management consoles, without using LAN in order to check OS starting situation, the LAN itself can be deleted. Thereby, in addition to the same effectiveness as the above-mentioned

second operation gestalt, further, the facility cost and the maintenance cost of LAN can be reduced and it is effective in low cost-ization being further realizable.
[0095] Moreover, in order according to this operation gestalt to communicate with a management console and to perform starting control, it is effective in the ability to reduce the load to storage further. Furthermore, it is effective in grasp of management / starting situation of a computer becoming easy by centralizing starting control.
[0096]
[Effect of the Invention] According to this invention, at the time of OS starting of two or more calculating machines, since the load to storage can be distributed in time, the superfluous access concentration to storage can be controlled in the computer system which shares storage. Therefore, the degradation of storage can be prevented and it is effective in the ability to shorten OS warm-up time in the whole system. Moreover, according to this invention, starting of two or more computers is attained without a manager or an operator spending the new time and effort and the time amount about starting of a computer, and it is effective in leading to reduction of management cost or labor costs.
[0097] Since starting can furthermore be performed only by the fiber channel according to this invention, without using LAN, it is effective in the ability to reduce the introductory cost generated by laying two kinds of cables, LAN at the time of system installation, and a fiber channel, and the employment cost for a maintenance.

[Translation done.]

Reference

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－249908
（P2001－249908A）
(43)公開日　平成13年９月14日(2001.9.14)

(51)Int.Cl.⁷		識別記号	FI		テーマコード（参考）
G06F	15/177	670	C06F	15/177	670A 5B011
		672			672B 5B014
	1/26			13/10	340A 5B045
	9/445			13/14	310Y 5B076
	13/10	340		1/00	334J

審査請求　未請求　請求項の数11　ＯＬ　（全 17 頁）　最終頁に続く

(21)出願番号　特願2000－67150(P2000－67150)
(22)出願日　平成12年３月７日(2000.3.7)

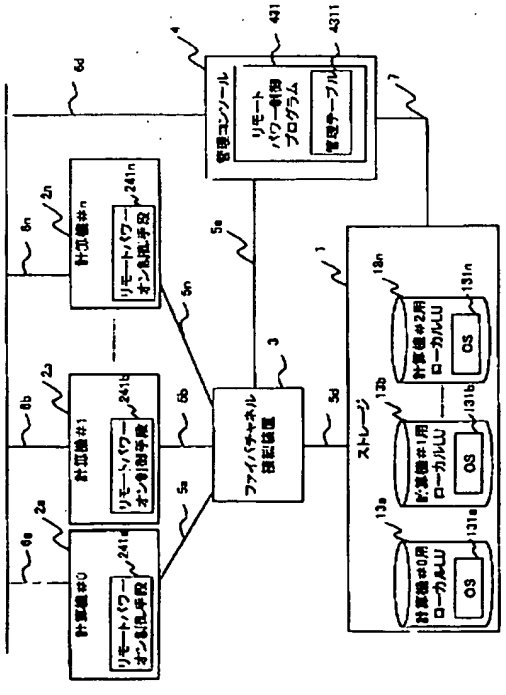
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　水野　陽一
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　松並　直人
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(74)代理人　100075096
弁理士　作田　康夫

最終頁に続く

(54)【発明の名称】　ストレージを共有する複数計算機のＯＳ起動の方法

(57)【要約】
【課題】複数の計算機がストレージを共有する計算機システムにおいて、従来、複数の計算機を同時に起動しようとするとストレージへのアクセスが集中し、ストレージの処理性能が低下し、OS起動に多大な時間がかかっていた。
【解決手段】少なくとも2台以上の複数の計算機と、前記計算機を起動するコンソールと、前記複数の計算機が共用するストレージを備え、ストレージは、前記計算機の各々が使用するOSを備え、コンソールは、前記計算機に対しパワーオンの指示を与えるリモートパワー制御プログラムと、前記計算機のグループ構成を管理する管理テーブルを備え、計算機は、遠隔地から電源制御できるリモートパワーオン制御手段を備え、前記リモートパワー制御プログラムは、前記管理テーブルのグループ構成に従い前記計算機をグループ毎にパワーオンする。

図 1



【特許請求の範囲】

【請求項１】　少なくとも2台以上の複数の計算機と、前記計算機を起動するためのコンソールと、前記複数の計算機が共用するストレージを備えた計算機システムにおいて、

ストレージは、前記計算機の各々が使用するOSを備え、

コンソールは、前記計算機に対しパワーオンの指示を与えるリモートパワー制御プログラムと、前記計算機のグループ構成を管理する管理テーブルを備え、

計算機は、遠隔地から電源制御をすることができるリモートパワーオン制御手段を備え、

前記リモートパワー制御プログラムは、前記管理テーブルのグループ構成に従い前記計算機をグループ毎にパワーオンすることを特徴とする計算機システム。

【請求項２】　前記請求項１記載の計算機システムにおいて、

前記リモートパワー制御プログラムが、前記管理テーブルのグループ構成に従い、前記計算機をグループ毎に時間差をつけてパワーオンし、計算機のOS起動を実行することを特徴とする計算機の起動の方法。

【請求項３】　少なくとも2台以上の複数の計算機と、前記全ての計算機が、ストレージ用インタフェースを介して直接アクセス可能なストレージと接続している計算機システムにおいて、

ストレージは、各々の計算機専用に設けたローカル論理ボリュームと、全ての計算機がアクセス可能な通信論理ボリュームを備え、

通信論理ボリュームは、全ての計算機の起動状況を記録する計算機ステータステーブルを備え、

計算機は、OSの起動を制御し計算機の起動状況を伝えるブートアップ制御プログラムを備え、

計算機ステータステーブルは、ブートアップ制御プログラムによって各々の計算機の起動状況が記載され、ブートアップ制御プログラムは、起動中の計算機数を知るために計算機ステータステーブルを参照し、その起動状況によって計算機のOS起動開始のタイミングを制御することで、ストレージへのアクセス分散を可能にし、計算機のOS起動を実行することを特徴とする計算機システム。

【請求項４】　前記請求項３記載の計算機システムにおいて、

前記計算機がパワーオンし、前記ブートアップ制御プログラムを始動するステップと、

前記ブートアップ制御プログラムが前記通信論理ボリュームの計算機ステータステーブルにアクセスし、全計算機の起動状況を参照するステップと、

前記ブートアップ制御プログラムが、前記計算機ステータステーブルの各々の計算機の専用領域に起動状況を記載するステップとを備えたことを特徴とする計算機の起動の方法。

【請求項５】　前記請求項３記載の計算機システムにおいて、

ストレージは、全ての計算機とコンソールがそれぞれアクセス可能な通信論理ボリュームと呼ぶ共用論理ボリュームを備え、

通信論理ボリュームは、各々の計算機専用に、その計算機だけが書き込み可能であると定めた専用論理領域を備え、

前記専用論理領域は、各々の計算機の起動状況を記載するものであり、前記計算機が各計算機の前記専用論理領域を適当なタイミングでリードすることにより、各計算機における起動状況を認識することを特徴とする通信の方法。

【請求項６】　前記請求項３記載の計算機システムの構成に、さらにコンソールを備えた計算機システムにおいて、

計算機は、OSの起動を制御し計算機の起動状況を伝えるブートアップ制御プログラムを備え、

コンソールは、計算機の起動を制御する計算機起動制御プログラムと計算機の起動状況を記録する計算機ステータステーブルを備え、

ブートアップ制御プログラムは計算機起動制御プログラムに起動状況を通信し、計算機起動制御プログラムは、計算機ステータステーブルを参照し、計算機の起動数を制御するためにブートアップ制御プログラムに通信することで、ストレージへのアクセスの時間的分散を可能にし、計算機のOS起動を実行することを特徴とする計算機システム。

【請求項７】　前記請求項６記載の計算機システムにおいて、

前記計算機がパワーオンし、前記ブートアップ制御プログラムを始動するステップと、前記ブートアップ制御プログラムが前記計算機起動制御プログラムに起動状況を通知するステップと、前記計算機起動制御プログラムが前記計算機ステータステーブルを参照するステップと、前記計算機起動制御プログラムが前記計算機ステータステーブルを書き換えるステップと、前記計算機起動制御プログラムが前記ブートアップ制御プログラムに通知するステップを備えたことを特徴とする計算機の起動の方法。

【請求項８】　前記請求項６記載の計算機システムにおいて、

前記コンソールおよび計算機が接続するためのインタフェースはファイバチャネルであり、

前記コンソールおよび計算機はファイバチャネルを制御する手段を備え、

前記コンソールおよび計算機が備えるファイバチャネルを制御する手段は、ブロードキャスト可能なメッセージフレームで、その内部に、少なくとも１台以上の計算機に実装したファイバチャネルを制御する手段を特定するための識別子であるWorld Wide Nameと、前記計算機の

起動状況や前記コンソールが発行する制御情報を特定するための起動制御識別子を格納したメッセージフレームを送信可能であり、
前記ファイバチャネルを制御する手段は、前記コンソールおよび計算機が発行したブロードキャストのメッセージフレームに、当該ファイバチャネルを制御する手段を特定するための識別子であるWorld Wide Nameが格納されていることを確認した時には、前記起動制御識別子の内容に従って起動制御を実行することを特徴とする計算機とコンソールの通信の方法。
【請求項９】　前記請求項３および６記載の計算機システムにおいて、
前記ブートアップ制御プログラムは、起動数制御により待機中の状態であることを前記計算機の操作者に伝えるために、前記計算機の出力装置にその旨を表示することを特徴とする計算機の起動の方法。
【請求項１０】　前記請求項３および６記載の計算機システムにおいて、
前記計算機が備えるブートアップ制御プログラムは電源起動時に実行されるプログラムであり、
前記計算機の起動数を制御し、前記計算機によって書き換え可能であるブートアップ制御プログラムを備えたことを特徴とする計算機システム。
【請求項１１】　前記請求項１から１０記載の計算機システムまたは計算機の起動の方法において、
ストレージ用インタフェースはファイバチャネルであることを特徴とする計算機システム及び計算機の起動の方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ストレージを共用する複数計算機のOS起動の方法に関し、特に、OS起動時に複数計算機のストレージ集中アクセスによる性能低下を防止する方法に関する。
【０００２】
【従来の技術】複数の計算機を用いた計算機システムとして、複数の計算機と、複数の計算機を管理するサーバと、複数の計算機が共用するストレージを備えた計算機システムが知られている。図１７の例を用いて説明する。
【０００３】同図において、200x（200a＝200n）は計算機、100は計算機200xが共用するストレージ、1001は計算機200xがダウンロードするOS、400は計算機を設置してある場所とは物理的に離れた場所にあるサーバ、500はサーバ400とストレージ100を接続する接続インターフェース、600x（600a＝600n）は複数の計算機200xとサーバ400が通信するためのLAN（Local Area Network）である。
【０００４】計算機200xがパワーオンすると、まず、計算機200x はLAN 600x経由でサーバ400にアクセスする。次に、ストレージ100にあるOS 1001をダウンロードし、起動を開始する。複数の計算機 200xがパワーオンすると、パワーオンした全ての計算機 200xがOS 1001をダウンロードしようとするため、サーバ 400やストレージ100への負荷が増大することになる。
【０００５】
【発明が解決しようとする課題】（課題）上記従来技術では、複数の計算機を同時に起動しようとすると、サーバやストレージにアクセスが集中することになり、OS起動のような負荷のかかる処理の場合、処理性能が低下し、OS起動に多大な時間がかかるという課題がある。同時に起動する計算機数が多いほどOS起動時間は長大化し、多台数の計算機を使用する場合、業務に支障を来すことになる。
【０００６】それぞれの操作者が電源投入時間をずらして起動すればストレージへのアクセス集中は回避できるが、他の操作者がいつ電源を入れるかを把握し、他の計算機の起動を待ち、常に電源投入時間を考慮しながら計算機に電源を投入しなけらばならないという課題がある。電源投入は計算機を使用するたびに行なわなければならず、操作者はその都度余計な手間がかかるという課題がある。
【０００７】（目的）本発明の目的は、ストレージを共用する計算機システムにおいて、管理者または操作者が、計算機の起動に関する新たな手間や時間を費やすことなく複数計算機の起動が可能な方法を提供し、かつ、複数計算機のOS起動時に、ストレージへの負荷を時間的に分散する方法を提供することにより、ストレージへのアクセス集中による性能低下を防止し、OS起動時間を短縮することである。
【０００８】
【課題を解決するための手段】上記の目的を達成するために、複数の計算機と、複数の計算機をパワーオンするための管理コンソールと、複数の計算機と管理コンソールが共用するストレージを備えた計算機システムを構築する。
【０００９】本発明の第一の手段では、管理コンソールは複数の計算機の電源を遠隔地から制御するためのリモートパワー制御プログラムと、計算機グループを管理するための管理テーブルを備える。
【００１０】計算機は、遠隔地から電源制御をすることができるリモートパワーオン制御手段を備える。
【００１１】リモートパワー制御プログラムは、管理テーブルで定義された計算機グループにしたがって計算機グループ毎に時間差をつけてパワーオンすることにより、ストレージへのアクセスの時間的分散を可能とする。
【００１２】本発明の第二の手段では、複数の計算機及び管理コンソールは、ファイバチャネル等の接続インターフェースによりストレージに接続する。

【００１３】管理コンソールは複数の計算機の電源を遠隔地から制御するためのリモートパワー制御プログラムを備える。

【００１４】計算機は、遠隔地から電源制御をすることができるリモートパワーオン制御手段と、OSの起動を制御し、計算機の起動状況を伝えるブートアップ制御プログラムを備える。

【００１５】ストレージは、複数の計算機のそれぞれ専用の論理ユニットと、すべての計算機が使用することができ、計算機の起動状況を記録する通信論理ユニットと、これらの論理ユニットを定義・作成するＬＵ定義プログラムとを備える。

【００１６】ブートアップ制御プログラムが通信論理ユニットを参照し、計算機数を制御することで、ストレージへのアクセス分散を可能にし、計算機のOS起動を実行する。

【００１７】また、本発明の第三の手段は、前記第二の手段において、ストレージに配置された通信論理ユニットのかわりに、管理コンソールに計算機の起動を制御する計算機起動制御プログラムを備え、集中型起動制御を行なうことを特徴とする。

【００１８】

【発明の実施の形態】（第一実施形態）従来の技術ではサーバを用いたシステム構成がとられていたが、本実施形態では、運用管理や性能の面で優れているサーバレス型のシステム構成とする。この場合も従来技術で述べたような複数計算機の起動時にストレージへのアクセス集中が起こるという課題は同様である。なお、もちろん本発明はこの構成に限定されるものではない。

【００１９】図１は本実施形態の計算機システムの構成図である。同図において、2x（2a＝2n）は計算機、1はすべての計算機2が共用するストレージ、4は計算機システムを管理するための管理コンソール、3はすべての計算機2xとストレージ1と管理コンソール4を相互に接続するファイバチャネル接続手段、5x（5a＝5n）はファイバチャネルである。7はストレージ1と管理コンソール 4が通信するための通信手段、6x（6a＝6n）は複数の計算機2xと管理コンソール4が通信するためのLAN（Local Area Network）である。

【００２０】以下、計算機としてはPC（Personal Computer）等のクライアント用計算機を想定し説明するが、サーバや、ワークステーション等のその他の計算機にも応用可能である。また、計算機とストレージを接続するためのインタフェース（以下Ｉ／Ｆと略記する）は、その接続距離と速度の面で最適であるファイバチャネルを想定するが、SCSI（Small Computer System Interface）、USB（Universal Serial Bus）、IEEE1394、等、様々なI／Fに適用可能である。

【００２１】図２は計算機2xの構成図である。21はこの計算機2のユーザが情報を入力したり出力したりするための入出力装置、22は計算機の全体の制御を行う中央制御手段、25はファイバチャネルの接続、制御を行うために計算機2に搭載したファイバチャネルボード、24はLAN 6x用いた通信を行うためのLANボード、23は中央制御手段22が各種制御を実行するために必要なプログラムやデータを格納するためのメモリである。

【００２２】ファイバチャネルボード25において、250はファイバチャネルの制御を行うファイバチャネルコントローラ、251はファイバチャネルコントローラが制御をするために必要なプログラムやデータや、中央制御手段22がファイバチャネルコントローラ250を制御するために必要なプログラムやデータを格納するためのメモリである。メモリ251において、2511はファイバチャネル5xに接続した計算機2xの外部のストレージ1からOSを起動するめにファイバチャネルコントローラ250を制御するために中央制御手段22が実行するブートアップ制御プログラムである。2512はファイバチャネルコントローラ250を一意に識別するために設けた世界で唯一な名称の情報であるWWN（World Wide Name）である。ここで、本実施形態ではファイバチャネルボード25が計算機2xに搭載されているとしたが、本ボードに搭載するファイバチャネルコントローラ250、およびメモリ251を計算機2xに該ボード25を用いずに直接実装することも可能であり、この場合も本発明の作用、効果は全く同一である。

【００２３】26は計算機2xのメイン電源スイッチである。

【００２４】図３はストレージ1の構成図である。11はストレージの全体制御を司る中央制御手段、15はファイバチャネル5dへの接続とその制御を行うファイバチャネルコントローラ、12は中央制御手段11がストレージ1の内部制御を行うためのプログラムやデータを格納するためのメモリ、16は管理コンソール4との間の通信を行うための通信手段、13x（13a＝13n）と14は論理ユニット（LU：Logical Unit）である。

【００２５】ここで論理ユニットLUとは、論理ボリュームとも呼ばれるストレージ1内部に設けた仮想的なボリュームのことであり、計算機とストレージを接続するI／Fの一つのプロトコルであるSCSI（Small Computer System Interface）の仕様において定義された名称である。以下、論理ボリュームのことを単にLUと呼ぶことにする。またLUを識別するための番号のことをLUN（Logical Unit Number）と呼ぶ。

【００２６】論理ユニット13xは計算機2x専用のLUであり計算機#x用ローカルLUである。ローカルLU13xには各計算機2x専用のOS 131xが格納される。一方、論理ユニット14はすべての計算機2xと管理コンソール4のいずれもがアクセス可能な共用LUである。

【００２７】メモリ12において、121は複数のLUを定義したり作成したりするために中央制御手段11が実行するLU定義プログラム、122は各計算機2xが複数のLUのそれ

ぞれに対するアクセス禁止／許可や、LUの属性や、上位計算機から認識されるLUNである仮想LUNとストレージ内部でLUを管理するために付加した内部LUNのそれぞれの対応関係を中央制御手段11が管理するためのLUN管理テーブル、123はLUN管理テーブル122の情報に従い各計算機2xからの各LUへのアクセスを制限／許可する制御を行い、許可した場合には当該LUに対するリード・ライトアクセスの制御を中央制御手段11が実行するためのアクセス制御プログラムである。

【００２８】LUについて説明する。LUはストレージの上位計算機から見たときの論理的なボリュームである。上位計算機は1つのLUを１台の論理的なディスク装置として認識する。

【００２９】ストレージ1は内部に複数のLUを定義し、構築することができる。これを内部LUと呼ぶことにする。ストレージ1では内部LUを管理するため0から始まる整数でシリアル番号付けする。この番号を内部LUNと呼ぶ。

【００３０】一方、本発明の計算機システムのように複数の計算機2xが１台のストレージ1を共用する場合、計算機2xそれぞれに専用のLUを割り当てる。一般にPC等の上位計算機はOSブート時に接続するストレージをサーチしてLUを検出するが、サーチ方法にいくつかの制約がある場合がある。それは、

(a) LUNは0から順にサーチする

(b) LUNは連続番号で存在することを仮定し、ある番号が存在しない場合以降のサーチは行わない

の２点である。これはサーチ時間を短縮するための工夫である。本発明の計算機2もこのような特性をもつ計算機2xであると仮定する。このような場合、もし、内部LUNをそのまま上位計算機に割り当てるとすると、内部LUN=0以外のLUを割り当てられた計算機はこのLUを検出できないことになってしまう。そこで、すべての計算機に対し、0から始まり、かつ連続番号でLUNを割り当てることが望ましい。そこで、本発明においては、ストレージ1は各計算機2xごとにその計算機2xが使用する内部LUををすべて0から始まる連続したLUNになるよう再定義することで上記の課題を解決する。このように各計算機2xから認識されるLUを仮想LU、その番号を仮想LUNと呼び、内部LUおよび内部LUNと区別する。これらの内部LUN、仮想LUNとの対応関係は、ストレージ1が備えるLUN管理テーブル122で管理する。

【００３１】ストレージ1が備えるLUN管理テーブル122の一例を図４に示す。LUN管理テーブル122には、ポート番号と、Target IDと、仮想LUNと、内部LUNと、WWNと、S_IDと、属性を格納する。

【００３２】ポート番号は、ストレージ1が備えるファイバチャネル接続ポートの番号を格納する。本実施形態ではポート数は１個を仮定し、一律0を格納する。

【００３３】TargetIDは、計算機2xとの接続I／Fにおいて、ストレージ1に割り当てる識別IDのことである。本実施形態のように計算機2xとストレージ1の接続I／Fがファイバチャネルの場合には各ポート毎に唯一のD_ID（Destination ID）を備えるが、ポート番号の項があるので省略してもよいし、ファイバチャネルの初期化時に決定したD_IDを格納しても良い。SCSIの場合には同一ポートに複数のIDを備えることができるので、そのときの各LUNの属するTarget IDを格納する。本実施形態ではファイバチャネルを仮定しているので、Target IDの欄は未使用とし、一律0を格納する。

【００３４】仮想LUNと内部LUNは、計算機2xに割り当てた内部LUNと各計算機2xから認識される仮想LUNとの対応関係を示す。たとえば、内部LUN0は仮想LUN0として、また内部LUN2も仮想LUN0として定義する。両者はともに仮想LUN0として定義されているが、それぞれ使用できる計算機2xが異なる。

【００３５】WWNは、各計算機2xのファイバチャネルコントローラ250を特定する情報であるWorld Wide Nameを格納する。ファイバチャネルのポートとポートの接続関係を確立するポートログイン処理の際に、各計算機2xのWWN 2512がストレージ1に通知される。

【００３６】S_IDは、ファイバチャネルのフレームヘッダに格納されるID情報であり、フレームを作成したソース（イニシエータ）を識別するIDである。S_IDは、ファイバチャネルの初期化の際に、動的に割り当てられる。先に述べたWWNは初期化の際に交換された各ファイバチャネルポートにより一意に設定される値であるが、WWNとS_IDの関連づけを行うことで、フレーム毎にWWNを調べなくてもS_IDのみ検査することで計算機2xが特定できるようになっている。

【００３７】属性は、各LUの所有属性を示す。「専用」は１台の計算機2x専用のLUであることを示す。「共用」は複数の計算機2xが共用するLUであることを示す。

【００３８】ここで、図３のストレージ1の構成図と図４のLUN管理テーブル122の対応関係について説明する。計算機2aがWWN＝WWNa、S_ID＝S_IDa、計算機2bがWWN＝WWNb、S_ID＝S_IDb、計算機2nがWWN＝WWNn、S_ID＝S_IDn、管理コンソール4がWWN＝WWNz、S_ID＝S_IDzとすると、図３の計算機2a用ローカルLU13aは内部LUN=0、仮想LUN=0、計算機2b用ローカルLU13bは内部LUN=1、仮想LUN=0、計算機2n用ローカルLU13nは内部LUN=n、仮想LUN=0、計算機2a、2b、・・・、2nの共用LU14は内部LUN=k、仮想LUN=1とマッピングする。また、共用LU14は管理コンソール4でも共用するので、WWN=WWNz、S_ID=S_IDzに対応する内部LUN=kは仮想LUN=0としてマッピングする。なお、ここで管理コンソール4はストレージ1に専用のローカルLUを所持していないので、共用LU14を仮想LUN=0としてマッピングしている。

【００３９】管理コンソール4を操作する作業者は、管理コンソール4の入出力装置41からLUを定義し、作成す

る指示を送ることができる。この指示は、通信手段47によりストレージ1に発行される。ストレージ1の通信手段16はこの指示を受信し中央制御手段11に報告する。中央制御手段11はLU定義プログラム121を起動し、指示に従いLUを作成する。そして、定義されたLUのアクセス制御情報として、ポート番号と、Target IDと、仮想LUNと、内部LUNと、WWNと、S_IDと、属性をLUN管理テーブル122に格納する。この作業をすべての計算機2xに対して割り当てるLUの回数繰り返すことで上記LUN管理テーブル122の通りLUを作成できる。

【００４０】図５は管理コンソール4の構成図である。41は管理者が管理コンソール4を操作するための入出力装置、42は管理コンソール4全体の制御を司る中央制御手段、43は中央制御手段41が実行するプログラムやデータを格納するためのメモリ、45はファイバチャネル5eへの接続、制御を行うファイバチャネルボード、44はLAN6dを用いた通信を行うためのLANボード、46は管理コンソール4のOSやプログラムを格納するディスク装置、47はストレージとの間の通信を行う通信手段である。

【００４１】メモリ43において、431は計算機2xの電源を制御するリモートパワー制御プログラム、4311は計算機グループを管理する管理テーブルである。図６に該テーブル4311の構成を示す。該テーブル4311は、各計算機のグループ番号が記載される。同じグループに属する計算機は同時に起動することになり、1グループあたりの計算機数は、ストレージの性能等によって決定する。

【００４２】同図は同時起動数を３としたときの例である。なお、該テーブル4311は一度作成しておけば、計算機構成の変更やストレージの処理性能の変更等がなければ、次回からもそのまま使用することができる。

【００４３】次に計算機の起動作業の詳細動作について説明する。

【００４４】(1)管理テーブル作成

管理コンソール4を操作する作業者は、起動する複数計算機をいくつかのグループに分割し、入出力装置41から各計算機のグループ番号を管理テーブル4311に登録する。起動しない計算機についてはグループ番号を０とする。

【００４５】(2)計算機のパワーオン

リモートパワー制御プログラム431は、管理テーブル4311にしたがって、グループ番号１から順に各グループ毎に一定間隔でリモートパワーオンパケット900の作成・送信を行い、計算機２ｘの起動を行なう。リモートパワーオンパケット900の送出間隔は、好ましくは計算機の起動時間と略同一とする。該パケット900は、計算機2xをパワーオンするためのパケットであり、その構成を図7に示す。該パケット900は、ヘッダ901とデータ部902からなる。データ部902は、同期部9021とリモートパワーオンする計算機2xのMACアドレス9022をｎ回連続したもので構成される。

【００４６】リモートパワー制御プログラム431はLANボード44を制御し同ボード44から該パケット900を計算機2xに送信する。

【００４７】計算機2xのLANボード24上にあるリモートパワーオン制御手段241は、該パケット900を受信する。同手段241は、該パケット900のデータ部902にあるMACアドレス9021を検出する。同手段241は、検出されたMACアドレス9021とLANボード24のMACアドレスとを比較し、同一であった場合はメイン電源スイッチ26を制御し、計算機2xをパワーオンする。パワーオンした計算機2xはストレージ1のローカルLU13xにある各計算機2x専用のOS 131xを読み込み、起動する。

【００４８】このようにして、計算機2xは、グループ毎に時間差をつけて順次起動していくので、ストレージへのアクセス集中の防止が可能となる。

【００４９】本実施形態によれば、計算機グループ毎に計算機の起動が行われ、常に一定数の計算機が起動するので、ストレージへのアクセスを時間的に分散することができ、過剰なアクセス集中を抑制することができる。したがって、ストレージの性能低下を防止することができ、システム全体でのOS起動時間を短縮することができる。

【００５０】（第二実施形態）第一実施形態では、管理コンソールによる遠隔からの計画的なパワーオンの場合には有効であるが、操作者による直接のパワーオンの場合には対応できないという課題が残る。したがって管理コンソールによるパワーオンと操作者によるパワーオンの両方が混在する環境においては、ストレージへアクセスが集中する可能性がある。パワーオンの方法が混在する環境においても対応できるように、パワーオンの方法にはよらないアクセス集中防止策が必要である。本実施形態は、パワーオンの方法にはよらずにストレージへのアクセス集中を防止し、複数計算機の起動を行なうことができる。ここでは、遠隔からの計算機のパワーオンはLAN経由であることを想定し、説明する。

【００５１】図８は本実施形態の計算機システムの構成図である。第一実施形態との相違点は、ストレージ 1に新たに通信LU 17を備えたことである。通信LU 17はすべての計算機2xと管理コンソール4のいずれもがアクセス可能であり、計算機の起動状況を通信するためのLUである。通信LU17には計算機ステータステーブル171を格納する。該テーブル171は計算機2xの起動状況を表すテーブルである。

【００５２】図９に該テーブル171の構成を示す。該テーブル171には排他制御を行なうためのフラグブロックと計算機2x用のブロックがあり、それぞれ該テーブル171内のアドレスを一意に決めておく。フラグブロックのステータス1はいずれかの計算機2xがテーブルにアクセス中であることを表し、ステータス0はテーブルにアクセス中の計算機がいないことを表す。計算機2x用ブロッ

クのステータス1はその計算機がOS起動中であることを表し、ステータス0は未起動、ステータス2は起動が終了したことを表す。計算機2xのブートアップ制御プログラム2511がファイルシステムを使用したファイルリード／ライトではなく、直接LUの論理ブロックをリード／ライトするRAW I／Oコマンドによりアクセスする。また、管理コンソール4は、この計算機ステータステーブル171をRAW I／Oでリードし、適当なタイミングでポーリングすることで各計算機2xの状況を逐次知ることができる。

【００５３】ブートアップ制御プログラム2511はファイバチャネル5xに接続した計算機2xの外部のストレージ1からOSを起動したり、通信LU117と通信するために用いられる。同プログラム2511は、フラッシュメモリ等を用いることにより、計算機2によって書き換え可能とすることもできる。

【００５４】次に計算機の起動作業の詳細動作について図１０のフローチャートを用いながら説明する。

【００５５】(a) 計算機のパワーオン

管理コンソール4のリモートパワー制御プログラム431による遠隔からのパワーオンや、操作者による直接のパワーオン等によって、計算機2xがパワーオンする（図１０のステップ701）。

【００５６】(b) ブートアップ制御プログラム始動

計算機2xはパワーオンするとファイバチャネルボード25のメモリ251にあるブートアップ制御プログラム2511を始動する（ステップ702）。

【００５７】(c) ブートアップ制御プログラムが計算機ステータステーブルを参照

ブートアップ制御プログラム2511は、ストレージ1の通信LU17にある計算機ステータステーブル171を参照する。

【００５８】該プログラム2511は、まずフラグブロックにアクセスし、アクセス中の計算機がいないか調べる。アクセス中の計算機がいなければフラグブロックを１に書き換え、計算機2x用のブロックを参照する。他の計算機がアクセス中であれば、アクセス終了まで待機する（ステップ703）。

【００５９】(d) 起動中の計算機数が閾値以下である

ブートアップ制御プログラム2511は、計算機ステータステーブル171の計算機2x用ブロックのステータス１の合計数によって起動中の計算機数を知ることができる。閾値はストレージの性能等によってあらかじめ決定された値であり、同時起動可能な計算機数を表す(ステップ704）。

【００６０】(e) 待機状態を画面表示

起動中の計算機数が閾値より多い場合は、フラグブロックのステータスを0に書き換え、一旦アクセスを終了し、他の計算機がOSの起動を終了して、起動中の計算機数が閾値以下になるまでブートアップ制御プログラム2511は待機する。その際、該プログラム2511は待機状態であることを操作者に伝えるため、入出力装置21の画面上に待機状態であることを表示する。該プログラム2511は一定時間経過した後、再び計算機ステータステーブル171にアクセスする（ステップ709）。

【００６１】(f) ブートアップ制御プログラムは自計算機のステータスを１に書き換える

起動中の計算機数が閾値以下であれば、ブートアップ制御プログラム2511は、次の処理に進む。起動中であることを表すため、該プログラム2511は、計算機ステータステーブル171の自計算機のステータスを1に書き換え、フラグブロックのステータスを0に戻し、該テーブル171へのアクセスを終了する（ステップ705）。

【００６２】(g) ブートアップ制御プログラムはOS起動開始

ブートアップ制御プログラム2511は、ストレージ1の計算機専用ローカルLU 13xにあるOS 131xの起動を開始する（ステップ706）。

【００６３】(h) OS起動完了後、ブートアップ制御プログラムは自計算機のステータスを0に書きかえる。

【００６４】OS 131xの起動が完了したことを表すため、ブートアップ制御プログラム2511は、計算機ステータステーブル171の自計算機のステータスを0に書き換える（ステップ707）。

【００６５】以上で計算機の起動動作が終了する。複数の計算機2xが同時に起動動作を実行した場合、閾値までは同時にOS 131xの起動処理が行なわれるが、閾値を超えた場合は、起動中の計算機2xがOS 131xの起動を完了し、計算機ステータステーブル171のステータスを書き換えるまで、他の計算機2xは待機することになる。これによって、ある一定数以上の計算機が同時にOS起動処理を行なうことがなくなり、ストレージ１への過剰なアクセス集中を抑制することができる。したがって、ストレージ1の性能低下を防止することができ、システム全体でのOS起動時間を短縮することができる。

【００６６】本実施形態によれば、複数計算機のOS起動時に、ストレージへの負荷を時間的に分散することができ、ストレージを共用する計算機システムにおいて、ストレージへの過剰なアクセス集中を抑制することができる。したがって、ストレージの性能低下を防止することができ、システム全体でのOS起動時間を短縮することができる。

【００６７】また、パワーオンの方法によらず適用可能であるので、遠隔からのパワーオンや直接のパワーオンが混在する環境においても有効である。

【００６８】また、管理者または操作者が、計算機の起動に関する新たな手間や時間を費やすことなく複数計算機の起動が可能となり、管理コストや人件費の削減につながる。

【００６９】（実施形態３）本実施形態においても管理コンソールによる遠隔からの計算機のパワーオンだけで

なく、操作者による直接のパワーオンの場合にも適用可能であり、パワーオンの方法にはよらずにストレージへのアクセスを分散した計算機の起動が可能である。ここでは、遠隔操作によるファイバチャネル経由での計算機のパワーオンを想定し、説明する。

【００７０】図１１は本実施形態の計算機システムの構成図である。第二実施形態との相違点は、図８のLAN 6が存在していないことと通信LU 17に配置されていた計算機ステータステーブル171を管理コンソール4に配置し、新規に計算機起動制御プログラム432を配置したことであり、その他は同一である。

【００７１】図１２は本実施形態の管理コンソールの構成図である。第二実施形態との相違点は、LANボードが存在しないことと、新規に計算機起動制御プログラム432を設け、ディスク装置46に計算機ステータステーブル171を格納したことである。計算機起動制御プログラム432は、OS起動処理をする計算機数を制御するためのプログラムである。計算機ステータステーブル171は図9に示すものと同一であり、計算機の起動状況を表すテーブルである。第二実施形態ではブートアップ制御プログラム2511が直接参照し、リード／ライトを行なったが、本実施形態では、ブートアップ制御プログラム2511は計算機起動制御プログラム432に通知し、計算機起動制御プログラム432がリード／ライトを行なう。

【００７２】図１３は本実施形態の計算機2xの構成図である。第二実施形態の計算機2xとの相違点は、LANボードが存在しないことと、ファイバチャネルボード上にリモートパワーオン制御手段2501と待機電源2502が存在していることである。

【００７３】待機電源2502は計算機2xに設けたものであり、計算機2xのメイン電源がオフの状態でも、この待機電源2502だけはオンになっているように構成する。

【００７４】リモートパワーオン制御手段2501は、ファイバチャネルコントローラ250の内部に設けたものであり、待機電源2502に接続している。よって、たとえ計算機2xのメイン電源がオフの状態でも、ファイバチャネルコントローラ250が動作するよう制御することができる。そこで、ファイバチャネル5経由で到着するフレーム（パケット）を受信することができる。

【００７５】次に計算機の起動作業の詳細動作について図１４のフローチャートを用いながら説明する。

【００７６】(a) 計算機のパワーオン

ファイバチャネルによるリモートパワーオン制御について説明する。図１５は、ファイバチャネルを用いたリモートパワーオンフレーム950の一実施形態である。9501はSOF（Start of Frame）であり、フレームの先頭を示す。9502はフレームヘッダであり、フレームの送信ノードの情報やフレームの種別等の情報を格納する。9502aはブロードキャスト識別子であり、このフレームがブロードキャストフレームとして、同一ファイバチャネルに接続するすべてのノードに対し同時に発行されるフレームであることを示す。9503はペイロードでありこのフレームの転送データの実態である。9503x（xはa=n）はWWNであり、リモートパワーオンを実施する計算機2xのファイバチャネルコントローラ250のWWN 2512を格納する。9504はCRC（Cyclic Redundancy Code）であり、フレームのビット化けを検出するためのデータ保証コードである。9505はEOF（End Of Frame）であり、フレームの末尾であることを示す。

【００７７】次にファイバチャネルを用いたリモートパワーオン制御の動作を説明する。管理コンソール4のリモートパワー制御プログラム4311は、リモートパワーオンフレーム950を作成する。このときリモートパワーオンの対象となる少なくとも１台以上の計算機2xのWWNをフレーム950のペイロード9503に格納する。そして、ファイバチャネルボード45を制御してブロードキャストで同フレーム950を送信する。

【００７８】ここで、各計算機2xはすべて停止していると仮定する。計算機2x本体は停止しているものの、ファイバチャネルコントローラ250は、待機電源2502で動作しており、ファイバチャネル5により到着するフレームを待ち受けている。ここで、上記のリモートパワーオンフレーム950を受信したとすると、同コントローラ250のリモートパワーオン制御手段2501はこのフレーム950のフレームヘッダ9502を検査し、ブロードキャストフレームであることを確認する。次に、同コントローラ250のWWN2512が同フレーム950のペイロード9503に格納されているかどうか検査する。もし、格納されていれば同手段2501はメイン電源スイッチ26を制御することで電源オンにして、計算機2xを起動する。もし、格納されていなければ、このフレーム950を無視し再び待ち受け状態になる。

【００７９】ここで、リモートパワーオンフレーム950のペイロード9503に格納したWWNは、第一実施形態に記載のLANによるリモートパワーオンパケットと同様に同一のWWNを繰り返し何回か格納しても構わない。

【００８０】以上のように、LANがなくてもファイバチャネル5を用いてリモートパワーオン制御が可能になる。

【００８１】上記においては、ファイバチャネルのブロードキャストフレームを用いパワーオン制御を行う方法を説明したが、もしファイバチャネルコントローラがIPプロトコルに対応しているならば、図7に示すリモートパワーオンパケットを用いる方法もある。この場合も上記と構成、動作は同様である。しかし、パケットをそのまま使用できるので、リモートパワー制御プログラムをそのまま適用できる。

【００８２】このようにして、複数計算機2xをパワーオンする（図１４のステップ801）。

【００８３】(b) ブートアップ制御プログラム始動

計算機2xはパワーオンするとファイバチャネルボード25のメモリ251にあるブートアップ制御プログラム2511を始動する（ステップ802）。

【００８４】(c) ブートアップ制御プログラムは管理コンソールの計算機起動制御プログラムに通知し、起動許可を待つ

ブートアップ制御プログラム2511は、起動開始であることを管理コンソール4の計算機起動制御プログラム432に通知するため、起動制御フレーム960を送出する。起動制御フレーム960は、ブートアップ制御プログラム2511が計算機2xの起動開始・終了を計算機起動制御プログラム432に通知したり、計算機起動制御プログラム432がブートアップ制御プログラム2511に起動許可を与えるために用いられるフレームである。図16に起動制御フレーム960の構成を示す。リモートパワーオンフレーム950との相違点は起動識別子96032が存在することである。計算機起動制御プログラム432は、WWN 96031によって計算機を特定し、起動識別子96032によってその計算機が起動開始か起動終了かを知ることができる（ステップ803）。

【００８５】(d) 計算機起動制御プログラムは、計算機ステータステーブルを参照

通知を受けた計算機起動制御プログラム432は、起動中の計算機数を知るために、ディスク装置46にある計算機ステータステーブル171を参照する。該テーブル171は図９に示すものと同一であり、計算機2xの起動状況を表すテーブルである。なお、本実施形態では該テーブル171を参照できるのは管理コンソールだけであり、排他制御が不要であるため、フラグブロックは使用しない（ステップ804）。

【００８６】(e) 起動中の計算機数が閾値以下である

計算機起動制御プログラム432は、計算機ステータステーブル171のステータス１の数によって起動中の計算機数を知ることができる。起動中の計算機数が閾値より多い場合は、他の計算機から起動終了通知が来て、起動中の計算機数が閾値以下になるまでブートアップ制御プログラム2511を待機させる。このとき、計算機起動制御プログラム432は、ブートアップ制御プログラム2511に待機命令を伝えるため起動制御フレーム960を送出する。起動中の計算機数が閾値以下であれば、計算機起動制御プログラム432は、次の処理に進む（ステップ805）。

【００８７】(f) 待機状態を画面表示

起動制御フレーム960が送出されると、ブートアップ制御プログラム2511はWWN96031によって自分宛のフレームかどうかを判断し、起動制御識別子96032によって待機命令を受けたことを知ることができる。待機命令を受けた該プログラム2511は、待機状態であることを操作者に伝えるため、入出力装置21の画面上に待機状態であることを表示する。該プログラム2511は一定時間経過した後、再び管理コンソール4の計算機起動制御プログラム432に起動開始の起動制御フレーム960を送出する（ステップ812）。

【００８８】(g) 計算機起動制御プログラムは通知を受けた計算機のステータスを１に書き換える

起動中の計算機数が閾値以下であれば、計算機起動制御プログラム432は、次の処理に進む。通知を受けた計算機が起動中であることを表すため、該プログラム432は、計算機ステータステーブル171の通知を受けた計算機のステータスを1に書き換える（ステップ806）。

【００８９】(h) 計算機起動制御プログラムは、ブートアップ制御プログラムに起動許可を与える

計算機起動制御プログラム432は、ブートアップ制御プログラム2511に起動許可を与えるため起動制御フレーム960を送出する。待機していたブートアップ制御プログラム2511はWWN 96031によって自分宛のフレームかどうかを判断し、起動制御識別子96032によって起動許可を受けたことを知ることができる（ステップ807）。

【００９０】(i) 起動許可を受けたブートアップ制御プログラムはOS起動開始

起動許可を受けたブートアップ制御プログラム2511は、ストレージ1の計算機専用ローカルLU13xにあるOS 131xの起動を開始する（ステップ808）。

【００９１】(j) OS起動完了後、ブートアップ制御プログラムは計算機起動制御プログラムに起動完了通知

ブートアップ制御プログラム2511は、計算機起動制御プログラム432に、OS 131xの起動が完了したことを通知するため、起動制御フレーム960を送出する（ステップ809）。

【００９２】(k) 計算機起動制御プログラムは通知を受けた計算機のステータスを0に書きかえる

OS 131xの起動が完了したことを表すため、計算機起動制御プログラム432は、計算機ステータステーブル171の通知を受けた計算機のステータスを0に書き換える（ステップ810）。

【００９３】以上で計算機の起動動作が終了する。複数の計算機2xが同時に起動動作を実行した場合、閾値までは同時にOS 131xの起動処理が行なわれるが、閾値を超えた場合は、起動中の計算機2xがOS 131xの起動を完了し、管理コンソールに起動終了を通知するまで、他の計算機2xは待機することになる。これによって、ある一定数以上の計算機が同時にOS起動処理を行なうことがなくなり、ストレージ１への過剰なアクセス集中を抑制することができる。したがって、ストレージ1の性能低下を防止することができ、システム全体でのOS起動時間を短縮することができる。第二実施形態が、各計算機による分散型起動数制御方式であったのに対し、本実施形態は、管理コンソールによる集中型起動数制御方式といえる。

【００９４】以上の通り、本実施形態によれば、OS起動状況を確認するために複数の計算機と管理コンソールの

間に必要な通信と、リモートパワーオン制御をLANを用いずにファイバチャネル経由で行うことができるので、LANそのものを削除することができる。これにより、上記第二実施形態と同一の効果に加え、さらに、LANの設備コストやメンテナンスコストを低減でき、さらに低コスト化を実現することができるという効果がある。

【００９５】また、本実施形態によれば、管理コンソールと通信して起動制御を行なうため、さらにストレージへの負荷を低減できるという効果がある。さらに、起動制御を集中化することにより、計算機の管理・起動状況の把握が容易になるという効果がある。

【００９６】

【発明の効果】本発明によれば、複数計算機のOS起動時に、ストレージへの負荷を時間的に分散することができるので、ストレージを共用する計算機システムにおいて、ストレージへの過剰なアクセス集中を抑制することができる。したがって、ストレージの性能低下を防止することができ、システム全体でのOS起動時間を短縮することができるという効果がある。また、本発明によれば、管理者または操作者が、計算機の起動に関する新たな手間や時間を費やすことなく複数計算機の起動が可能となり、管理コストや人件費の削減につながるという効果がある。

【００９７】さらに本発明によれば、LANを用いずにファイバチャネルだけで起動作業を実行できるので、システム導入時のLANとファイバチャネルの２種類のケーブルを敷設することにより発生する導入コストや、維持管理のための運用コストを低減することができるという効果がある。

【図面の簡単な説明】

【図１】第１実施形態の計算機システムの構成図。

【図２】第１実施形態の計算機の構成図。

【図３】第１実施形態のストレージの構成図。

【図４】第１実施形態のストレージのLUN管理テーブルの構成図。

【図５】第１実施形態の管理コンソールの構成図。

【図６】第１実施形態の管理テーブルの構成図。

【図７】第１実施形態のリモートパワーオンパケットの構成図。

【図８】第２実施形態の計算機システムの構成図。

【図９】第２実施形態の計算機ステータステーブルの構成図。

【図１０】第２実施形態の計算機の起動動作のフローチャート。

【図１１】第３実施形態の計算機システムの構成図。

【図１２】第３実施形態の管理コンソールの構成図。

【図１３】第３実施形態の計算機の構成図。

【図１４】第３実施形態の計算機の起動動作のフローチャート。

【図１５】第３実施形態のファイバチャネルのリモートパワーオンフレームの構成図。

【図１６】第３実施形態の起動制御フレームの構成図。

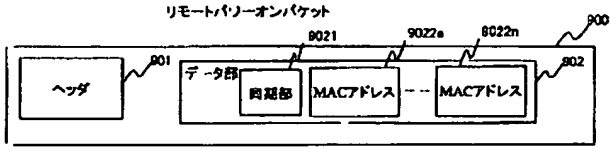
【図１７】従来技術の計算機システムの構成図。

【符号の説明】

1…ストレージ、11…中央制御手段、12…メモリ、13…計算機専用ローカルLU、14…共用LU、15…ファイバチャネルコントローラ、16…通信手段、17…通信LU、121…LU定義プログラム、122…LUN管理テーブル、123…アクセス制御プログラム、131…OS、171…計算機ステータステーブル、2…計算機、21…入出力装置、22…中央制御手段、23…メモリ、24…LANボード、241…リモートパワーオン制御手段、25…ファイバチャネルボード、250…ファイバチャネルコントローラ、2501…リモートパワーオン制御手段、2502…待機電源、251…メモリ、2511…ブートアップ制御プログラム、2512…WWN、26…メイン電源スイッチ、3…ファイバチャネル接続装置、4…管理コンソール、41…入出力装置、42…中央制御手段、43…メモリ、431…リモートパワー制御プログラム、4311…管理テーブル、432…計算機起動制御プログラム、44…LANボード、45…ファイバチャネルボード、46…ディスク装置、47…通信手段、5…ファイバチャネル、6…LAN、7…通信I／F、900…リモートパワーオンパケット、950…リモートパワーオンフレーム、960…起動制御フレーム、100…ストレージ、1001…OS、200…計算機、400…サーバ、500…接続インターフェース、600…LAN。

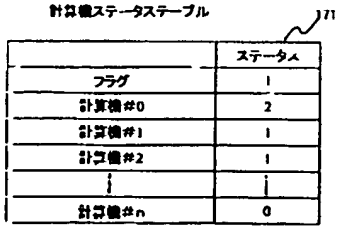
【図７】

図 7



【図９】

図 9



【図1】

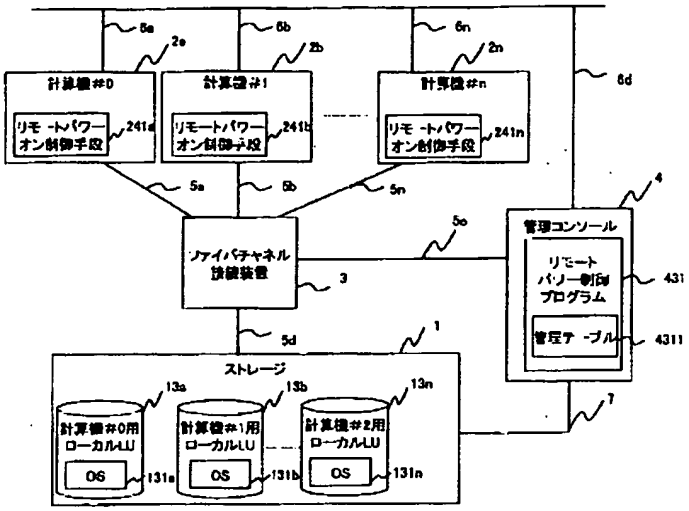


図1

【図6】

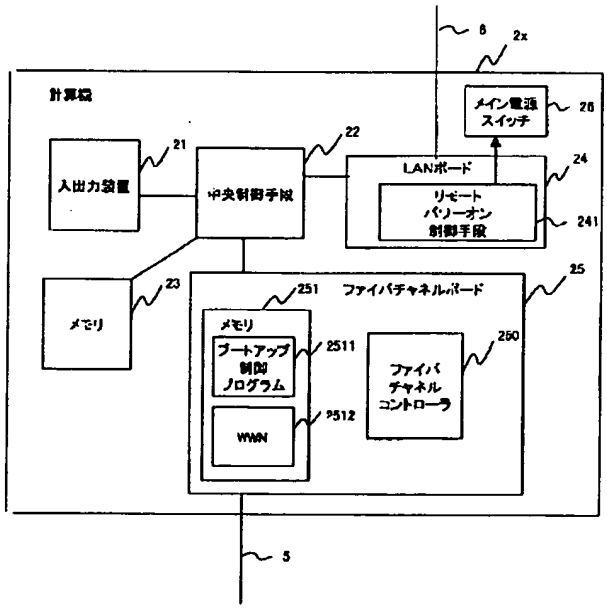
図 6

管理テーブル 4311

	グループ#
計算機#0	1
計算機#1	1
計算機#2	1
計算機#3	2
計算機#4	2
計算機#5	2
計算機#n	M

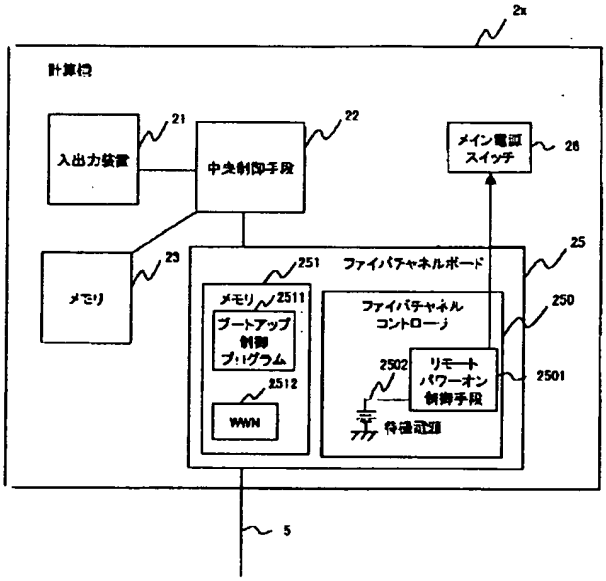
【図2】

図 2



【図13】

図 13



【図15】

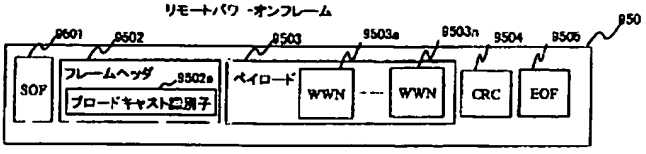


図15

【図3】

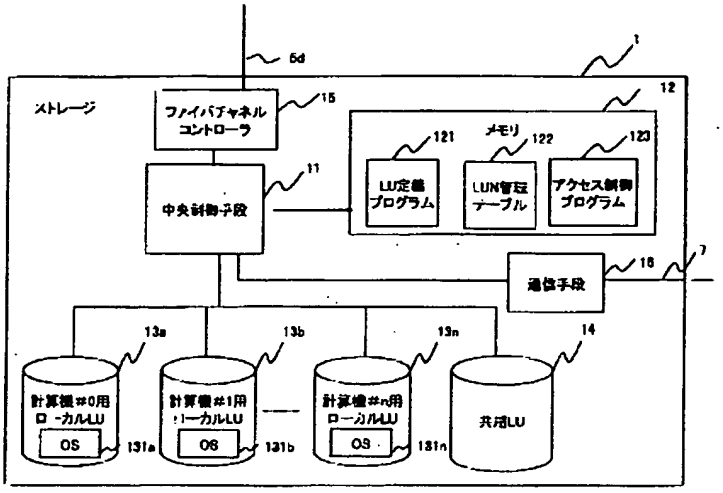


図3

【図4】

22

ポート番号	TargetID	仮想LUN	内部LUN	WWN	S_ID	属性
0	0	0	0	WWNa	S_IDa	専用
0	0	0	1	WWNb	S_IDb	専用
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	2	WWNc	S_IDc	専用
0	0	1	k	WWNa	S_IDa	共用
0	0	1	k	WWNb	S_IDb	共用
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	1	k	WWNn	S_IDn	共用
0	0	0	k	WWNz	S_IDz	共用

LUN管理テーブル

図4

【図5】

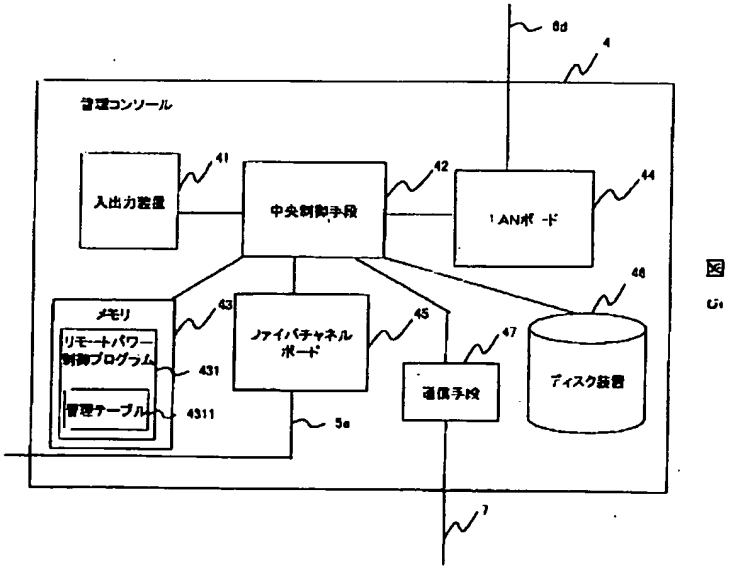


図5

【図8】

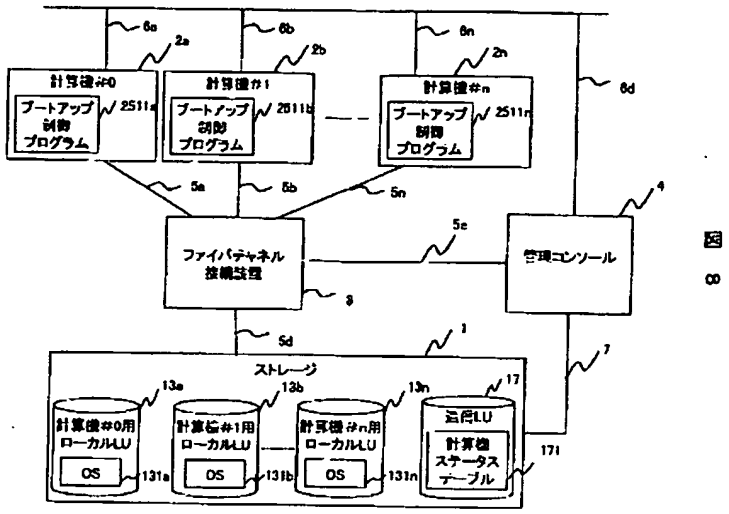


図8

【図11】

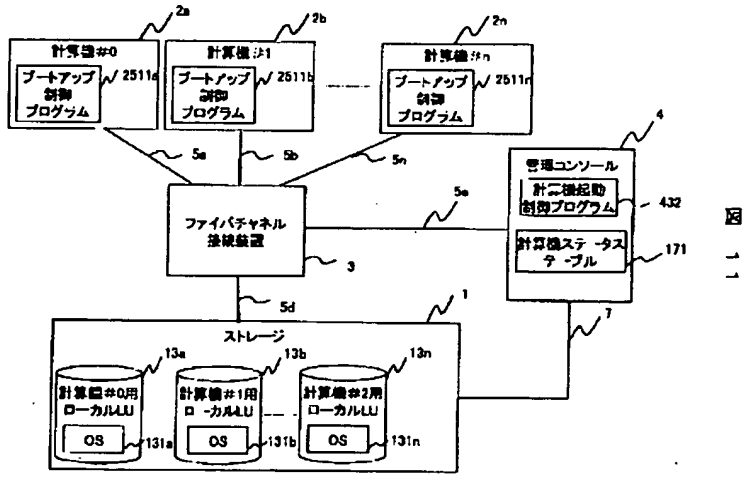


図11

【図16】

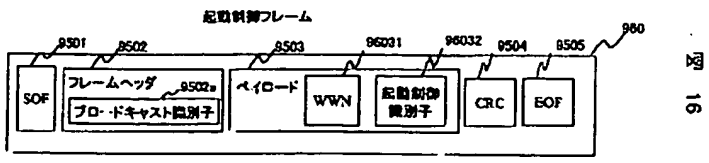
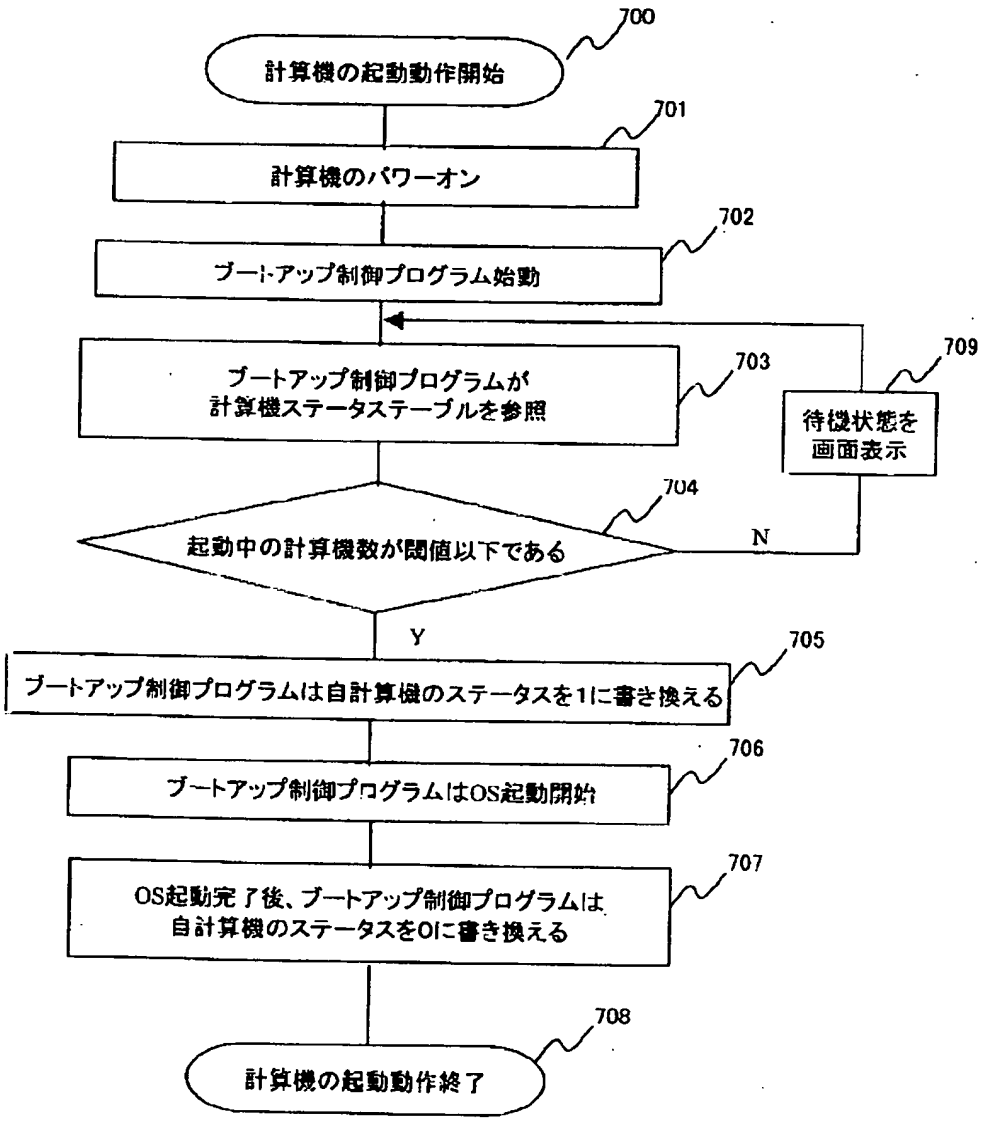


図16

【図10】

図　10


700
計算機の起動動作開始
701
計算機のパワーオン
702
ブートアップ制御プログラム始動
703
ブートアップ制御プログラムが
計算機ステータステーブルを参照
709
待機状態を
画面表示
704
起動中の計算機数が閾値以下である
N
Y
705
ブートアップ制御プログラムは自計算機のステータスを1に書き換える
706
ブートアップ制御プログラムはOS起動開始
707
OS起動完了後、ブートアップ制御プログラムは
自計算機のステータスを0に書き換える
708
計算機の起動動作終了

【図12】

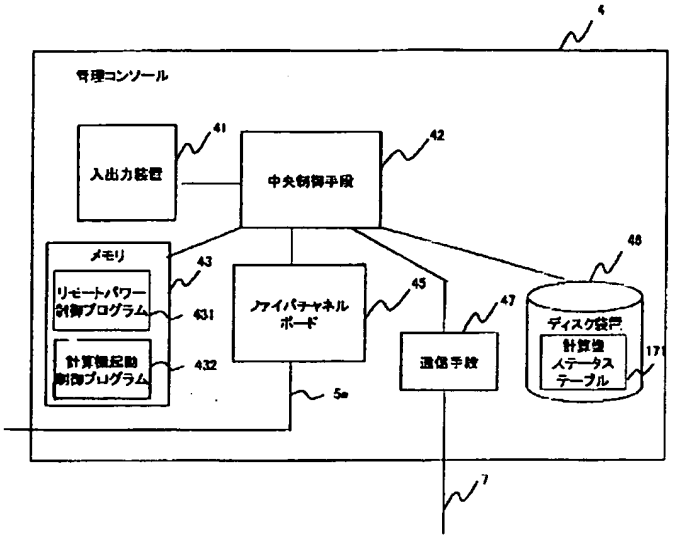


図12

【図17】

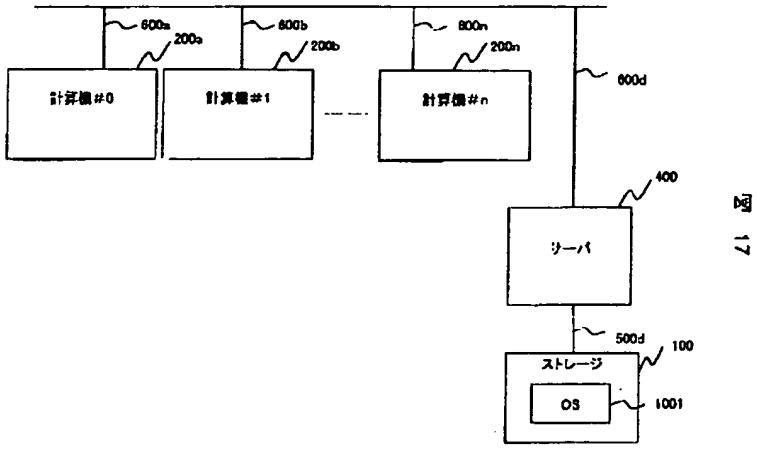
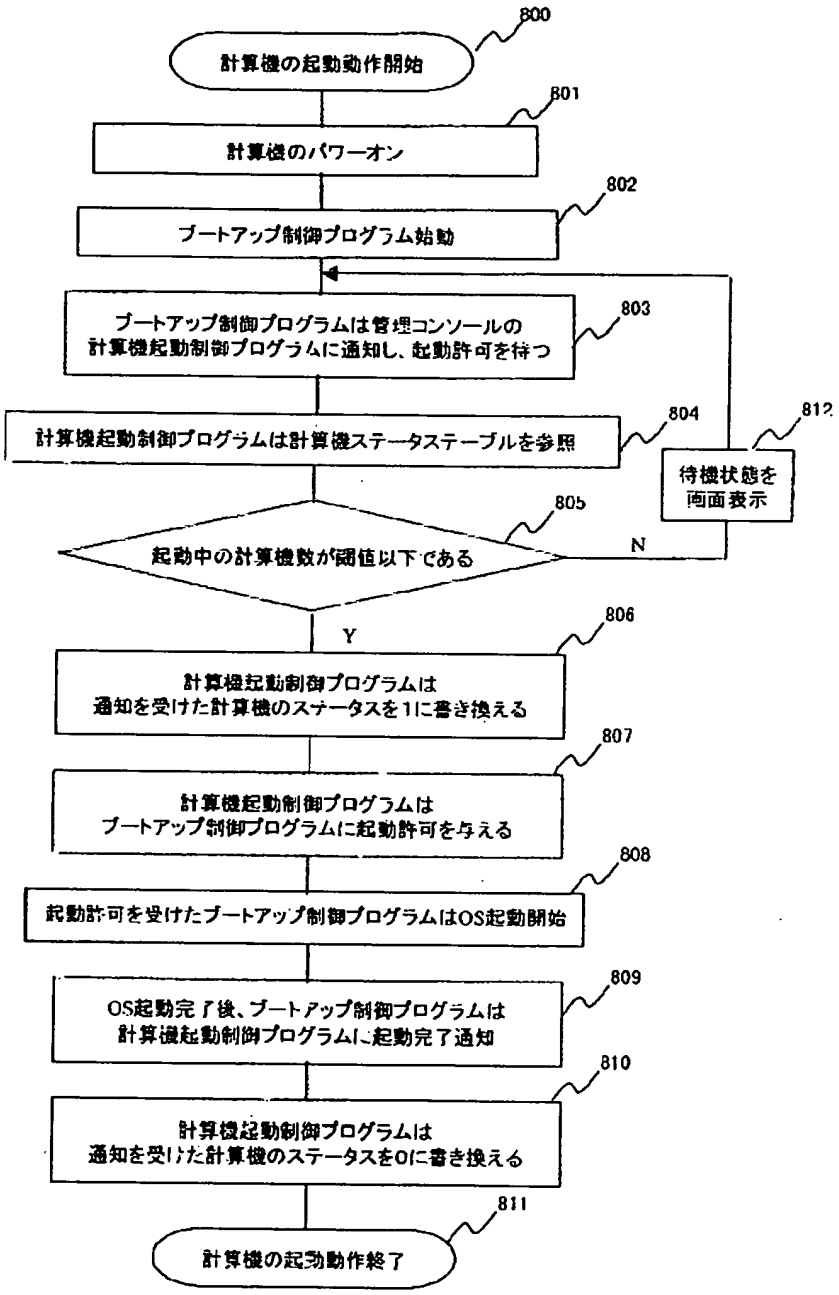


図17

【図14】

図　14



フロントページの続き

(51)Int.Cl.⁷	識別記号	FI		(参考)
G06F 13/14	310	G06F 9/06	420K	

(72)発明者　神牧　秀樹
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　松本　純
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

Fターム(参考)　5B011 EB07 MA04 MB06
5B014 EB05 FA04 HA02 HA07 HC17
5B045 DD04 HH01 KK03
5B076 AA13 BB02 BB06